¡¡APRUEBE SU EXAMEN CON SCHAUM!!

Estadística

Schaum

4ª EDICIÓN

Murray R. Spiegel • Larry J. Stephens

- 486 PROBLEMAS RESUELTOS PASO A PASO
- 660 PROBLEMAS DE PRÁCTICA
- INCLUYE SOLUCIONES A PROBLEMAS DE LOS PROGRAMAS DE CÓMPUTO ACTUALES

Mc Graw Hill

Utilícelo para las siguientes asignaturas:

☑ ESTADÍSTICA BÁSICA ☑ MÉTODOS ESTADÍSTICOS BÁSICOS ☑ INTRODUCCIÓN A LA ESTADÍSTICA
☑ INTRODUCCIÓN A LA PROBABILIDAD Y ESTADÍSTICA ☑ INTRODUCCIÓN AL ANÁLISIS EXPLORATORIO DE DAT

Estadística

ESTADÍSTICA

Cuarta edición

Murray R. Spiegel
Rensselaer Polytechnic Institute
Hartford Graduate Center

Larry J. Stephens
University of Nebraska at Omaha

Revisión técnica

Raúl Gómez Castillo
Instituto Tecnológico y de Estudios Superiores de Monterrey,
Campus Estado de México

MÉXICO • BOGOTÁ • BUENOS AIRES • CARACAS • GUATEMALA
LISBOA • MADRID • NUEVA YORK • SAN JUAN • SANTIAGO
AUCKLAND • LONDRES • MILÁN • MONTREAL • NUEVA DELHI
SAN FRANCISCO • SINGAPUR • SAN LUIS • SIDNEY • TORONTO

Director Higher Education: Miguel Ángel Toledo Castellanos
Director editorial: Ricardo A. del Bosque Alayón
Coordinadora editorial: Marcela I. Rocha Martínez
Editor sponsor: Pablo E. Roig Vázquez
Editora de desarrollo: Ana L. Delgado Rodríguez
Supervisor de producción: Zeferino García García
Traducción: María del Carmen Enriqueta Hano Roa

ESTADÍSTICA
Cuarta edición

Prohibida la reproducción total o parcial de esta obra,
por cualquier medio, sin la autorización escrita del editor.

DERECHOS RESERVADOS © 2009, respecto a la cuarta edición en español por
McGRAW-HILL/INTERAMERICANA EDITORES, S.A. de C.V.
A Subsidiary of ***The McGraw-Hill*** *Companies, Inc.*
Edificio Punta Santa Fe
Prolongación Paseo de la Reforma 1015, Torre A
Piso 17, Colonia Desarrollo Santa Fe
Delegación Álvaro Obregón
C.P. 01376, México, D. F.
Miembro de la Cámara Nacional de la Industria Editorial Mexicana, Reg. Núm. 736

ISBN-13: 978-970-10-6887-8
ISBN-10: 970-10-6887-8
(ISBN 970-10-3271-3 anterior)

Traducido de la cuarta edición de: *Theory and Problems of Statistics*.
Copyright © MMVIII by The McGraw-Hill Companies, Inc.
All rights reserved
ISBN: 978-0-07-148584-5

1234567890 0876543219

Impreso en México *Printed in Mexico*

The McGraw-Hill Companies

A la memoria de mi madre y de mi padre, Rosie y Johnie Stephens

L. J. S.

ACERCA DE LOS AUTORES

MURRAY R. SPIEGEL† obtuvo su maestría en física y su doctorado en matemáticas, ambos en Cornell University. Trabajó en Harvard University, Columbia University, Oak Ridge and Rensselaer Polytecnic Institute; además fue asesor en matemáticas para diversas compañías. Su último cargo fue como profesor y presidente de matemáticas en el Hartford Graduate Center del Rensselaer Polytecnic Institute. Su interés por las matemáticas lo acompañó durante toda su trayectoria, en especial en la rama que comprende la aplicación de la física y los problemas de ingeniería. Fue autor de numerosos artículos periodísticos y de más de una docena de libros sobre temas matemáticos.

LARRY J. STEPHENS es profesor de matemáticas en University of Nebraska at Omaha, donde imparte cátedras desde 1974. Su labor como docente la ha desarrollado también en instituciones como University of Arizona, Gonzaga University y Oklahoma State University. En su experiencia laboral destacan sus trabajos para la NASA, el Livermore Radiation Laboratory y el Los Alamos Laboratory. Desde 1989, el doctor Stephens es consultor e instructor en seminarios de estadística para grupos de ingeniería en 3M, en la planta de Nebraska. Ha colaborado en más de cuarenta publicaciones a nivel profesional. Es autor de numerosos bancos de pruebas computarizados, además de textos elementales de estadística.

PREFACIO A LA CUARTA EDICIÓN

Esta nueva edición contiene ejemplos nuevos, 130 figuras nuevas y resultados obtenidos empleando cinco paquetes de software representativos de los cientos o quizá miles de paquetes de software usados en estadística. Todas las figuras de la tercera edición han sido sustituidas por figuras nuevas, un poco diferentes, creadas empleando estos cinco paquetes de software: EXCEL, MINITAB, SAS, SPSS y STATISTIX. Los ejemplos tienen una gran influencia de *USA Today*, pues este periódico es una gran fuente de temas y ejemplos actuales de la estadística.

Otros de los cambios que se encontrarán en esta edición son: el capítulo 18 sobre análisis de serie de tiempos fue eliminado y el capítulo 19 sobre control estadístico de procesos y capacidad de procesos se convirtieron en el capítulo 18. Las respuestas a los ejercicios complementarios, al final de cada capítulo, se presentan ahora con más detalle. En todo el libro se analizan y emplean más los valores *p*.

RECONOCIMIENTOS

Dado que el software para estadística es muy importante en este libro, quiero agradecer a las personas y empresas siguientes por permitirme usar su software.

MINITAB: Laura Brown, coordinadora del Programa de Ayuda a los Autores, Minitab, Inc., 1829 Pine Hall Road, State College, PA 16801. Yo soy miembro del programa de ayuda a los autores que practica Minitab, Inc. "Partes de los datos y de los resultados que se encuentran en esta publicación/libro han sido impresas con la autorización de Minitab, Inc. Todo este material, así como los derechos de autor, son propiedad exclusiva de Minitab, Inc." La dirección de Minitab en la red es www.minitab.com.

SAS: Sandy Varner, directora de operaciones de mercadotecnia, SAS Publishing, Cary, NC. "Creado con el software SAS. Copyright 2006. SAS Institute Inc., Cary, NC." Se cita de su sitio en la red: "SAS es el líder en servicios y software inteligente para negocios. A lo largo de sus 30 años, SAS ha crecido —de siete empleados a casi 10 000 en todo el mundo, de unos cuantos clientes a más de 40 000— y todos estos años ha sido rentable". La dirección en la red de SAS es www.sas.com.

SPSS: Jill Rietema, gerente de cuenta, Publicaciones, SPSS. Se cita de su sitio en la red: "SPSS Inc. es líder como proveedor mundial de soluciones y software para análisis predictivo. Fundada en 1968, actualmente SPSS tiene más de 250 000 clientes en todo el mundo, atendidos por más de 1 200 empleados en 60 países." La dirección en la Red de SPSS es www.spss.com.

STATISTIX: Dr. Gerard Nimis, presidente, Analytical Software, P.O. Box (apartado postal) 12185, Tallahassee, FL 32317. Se toma de su sitio en la red: "Si se tiene que analizar datos y se es un investigador, pero no un especialista en estadística, STATISTIX está diseñado para ello. No necesitará programar ni usar un manual. Este software fácil de aprender y de usar ahorrará valioso tiempo y dinero. STATISTIX combina, en un solo y económico paquete, la esta-

dística, tanto básica como avanzada, con las poderosas herramientas para la manipulación de datos que se necesitan." La dirección en la Red de Statistix es www.statistix.com.

EXCEL: Se cuenta con Excel, de Microsoft, desde 1985. Cuentan con él casi todos los estudiantes universitarios. En este libro se emplea ampliamente.

Deseo dar las gracias a Stanley Wileman por la asesoría informática desinteresada que me proporcionó en la creación de este libro. Quiero agradecer a mi esposa, Lana, por su comprensión durante los días que dediqué a pensar en la mejor manera de presentar algunos conceptos. Mi agradecimiento a Chuck Wall, Senior Adquisitions Editor, y a su equipo de McGraw-Hill. Por último, quiero dar las gracias a Jeremy Toynbee, director de proyecto en Keyword Publishing Services Ltd., Londres, Inglaterra, y a John Omiston, copy editor independiente, por su excelente trabajo de producción.

LARRY J. STEPHENS

PREFACIO A LA TERCERA EDICIÓN

Al preparar esta tercera edición de Estadística, Serie Schaum, he reemplazado problemas antiguos por problemas que reflejan los cambios tecnológicos y sociológicos ocurridos desde que se publicó la primera edición en 1961. Por ejemplo, uno de los problemas en la segunda edición trata del tiempo de vida de los bulbos de radio. Como la mayoría de las personas menores de treinta años probablemente no sepan lo que es un bulbo de radio, este problema, lo mismo que muchos otros, fue sustituido por ejercicios que se refieren a temas actuales como el cuidado de la salud, el sida, Intenet, los teléfonos celulares, entre otros. Los asuntos matemáticos y estadísticos no han cambiado, sólo lo hicieron las áreas de aplicación y los aspectos de cálculo en estadística.

Otra mejora es la introducción en el texto de software para estadística. El desarrollo de software para estadística, como SAS, SPSS y Minitab, ha variado drásticamente las aplicaciones de la estadística a problemas de la vida real. El software para estadística más utilizado, tanto en el medio académico como en el industrial, es el Minitab. Quiero agradecer a Minitab Inc., por haberme otorgado el permiso para incluir, a lo largo de todo el libro, los resultados de Minitab. Muchos de los textos modernos de estadística traen, como parte del libro, resultados de algún paquete de software para estadística. En esta obra decidí emplear Minitab, ya que es muy utilizado y porque es muy amigable.

Una vez que el estudiante aprende las diversas estructuras de archivos de datos necesarios para utilizar Minitab, así como la estructura de comandos y subcomandos, puede transferir con facilidad ese conocimiento a otros paquetes de software para estadística. Gracias a la introducción de menús como las cajas de diálogo, el software resulta muy amigable. La obra adiciona tanto los menús como las cajas de diálogo que presenta Minitab. En muchos de los problemas nuevos se discute el importante concepto de pruebas estadísticas. Cuando se publicó la primera edición, en 1961, el valor p no se utilizaba tan ampliamente como ahora, debido a que con frecuencia resulta difícil determinarlo sin la ayuda de un software. En la actualidad, el software para estadística da el valor p de manera rutinaria, puesto que, con este apoyo, su cálculo es a menudo un asunto trivial.

Un nuevo capítulo titulado "Control estadístico de procesos y capacidad de procesos" reemplazó al capítulo 19, "Números índices". Estos temas tienen gran aplicación industrial, por lo que se agregaron al libro. La inclusión, en los paquetes de software modernos, de técnicas de control estadístico de procesos y capacidad de procesos ha facilitado su utilización en nuevos campos industriales. El software lleva a cabo todos los cálculos, que son bastante laboriosos.

Quiero agradecer a mi esposa Lana por su comprensión durante la preparación de este libro; a mi amigo Stanley Wileman, por la ayuda computacional que me brindó; y a Alan Hunt y su equipo de Keyword Publishing Service, en Londres, por su minusioso trabajo de producción. Por último quiero agradecer al equipo de McGraw-Hill por su cooperación y ayuda.

LARRY J. STEPHENS

PREFACIO A LA SEGUNDA EDIBCIÓN

La estadística, o los métodos estadísticos, como se llaman algunas veces, desempeñan un papel cada vez más importante en casi todas las áreas del quehacer humano. Aunque en un principio tenía que ver solamente con asuntos de Estado, a lo que debe su nombre, en la actualidad la influencia de la estadística se ha extendido a la agricultura, la biología, el comercio, la química, la comunicación, la economía, la educación, la electrónica, la medicina, la física, las ciencias políticas, la psicología, la sociología y a muchos otros campos de la ciencia y la ingeniería.

El propósito de esta obra es presentar una introducción a los principios generales de la estadística, que será útil a todos los individuos sin importar su campo de especialización. Se diseñó para usarse ya sea como consulta para todos los textos estándar modernos o como un libro para un curso formal de estadística. Será también de gran valor como referencia para todos aquellos que estén aplicando la estadística en su campo de investigación particular.

Cada capítulo empieza con una presentación clara de las definiciones correspondientes, los teoremas y principios, junto con algunos materiales ilustrativos y descriptivos. A esto le sigue un conjunto de problemas resueltos y complementarios, que en muchos casos usan datos de situaciones estadísticas reales. Los problemas resueltos sirven para ilustrar y ampliar la teoría, hacen énfasis en aquellos pequeños puntos importantes sin los cuales el estudiante se sentiría continuamente inseguro; además, proporciona una repetición de los principios básicos, aspecto que es vital para una enseñanza eficiente. En los problemas resueltos se incluyen numerosas deducciones de fórmulas. La cantidad de problemas complementarios con respuestas constituyen una revisión completa del material de cada capítulo.

Los únicos conocimientos matemáticos necesarios para la comprensión de todo el libro son la aritmética y el álgebra elemental. En el capítulo 1 viene una revisión de los conceptos matemáticos importantes, que se pueden leer al principio del curso o después, cuando la necesidad se presente.

Los primeros capítulos se ocupan del análisis de las distribuciones de frecuencia y de las correspondientes medidas de tendencia central, dispersión, sesgo y curtosis. Lo anterior lleva, de manera natural, a una discusión de la teoría de probabilidad elemental y sus aplicaciones, lo que prepara el camino para el estudio de la teoría del muestreo. De entrada, se abordan las técnicas de las muestras grandes, que comprenden la distribución normal, así como las aplicaciones a la estimación estadística y las pruebas de hipótesis y de significancia. La teoría de las muestras pequeñas, que comprende la distribución t de Student, la distribución ji cuadrada y la distribución F, junto con sus aplicaciones, aparecen en un capítulo posterior. Otro capítulo sobre ajuste de curvas y el método de mínimos cuadrados lleva, de manera lógica, a los temas de correlación y regresión que involucran dos variables. La correlación múltiple y la parcial, que involucran más de dos variables, son tratadas en un capítulo aparte. A este tema le siguen capítulos sobre el análisis de varianza y métodos no paramétricos, que son nuevos en esta segunda edición. Dos capítulos finales tratan de series de tiempo y número índice, en ese orden. Además, se ha incluido más material del que se alcanza a cubrir en un primer curso. El objetivo es hacer el libro más flexible para proporcionar una obra de referencia más útil y estimular un posterior interés en estos temas. La obra permite cambiar el orden de muchos de los últimos capítulos u omitir algunos sin dificultad. Por ejemplo, los capítulos 13 a 15 y 18 y 19 pueden ser introducidos, en su mayor parte, inmediatamente después del capítulo 5, si se desea tratar correlación, regresión, series de tiempo y números índice antes de la teoría del muestreo. De igual manera, dejar de lado la mayor parte del capítulo 6, si no se desea dedicar mucho tiempo a probabilidad. En un primer curso, en ocasiones el capítulo 15 se ignora en su totalidad. El orden se plantea debido a que en

los cursos modernos hay una tendencia creciente a introducir teoría del muestreo y la inferencia estadística tan pronto como sea posible.

Quiero agradecer a varias instituciones, tanto públicas como privadas, su cooperación al proporcionar datos para tablas. A lo largo del libro se dan las referencias apropiadas para esas fuentes. En particular, agradezco al profesor sir Roland A. Fisher, F.R.S., Cambrige; al doctor Frank Yates, F.R.S., Rothamster; y a Messrs. Oliver and Bond Ltd., Ediburgh, por haber otorgado el permiso para utilizar los datos de la tabla III de su libro *Statistical Tables for Biological, Agricultural, and Medical Research*. También quiero agradecer a Esther y a Meyer Scher, su apoyo, y al equipo de McGraw-Hill, su cooperación.

CONTENIDO

CAPÍTULO 1 Variables y gráficas 1

Estadística 1
Población y muestra; estadística inductiva (o inferencial) y estadística descriptiva 1
Variables: discretas y continuas 1
Redondeo de cantidades numéricas 2
Notación científica 2
Cifras significativas 3
Cálculos 3
Funciones 4
Coordenadas rectangulares 4
Gráficas 4
Ecuaciones 5
Desigualdades 5
Logaritmos 6
Propiedades de los logaritmos 7
Ecuaciones logarítmicas 7

CAPÍTULO 2 Distribuciones de frecuencia 37

Datos en bruto 37
Ordenaciones 37
Distribuciones de frecuencia 37
Intervalos de clase y límites de clase 38
Fronteras de clase 38
Tamaño o amplitud de un intervalo de clase 38
La marca de clase 38
Reglas generales para formar una distribución de frecuencia 38
Histogramas y polígonos de frecuencia 39
Distribuciones de frecuencia relativa 39
Distribuciones de frecuencia acumulada y ojivas 40
Distribuciones de frecuencia acumulada relativa y ojivas porcentuales 40
Curvas de frecuencia y ojivas suavizadas 41
Tipos de curvas de frecuencia 41

CAPÍTULO 3 Media, mediana y moda, y otras medidas de tendencia central 61
Índices o subíndices 61
Sumatoria 61
Promedios o medidas de tendencia central 62
La media aritmética 62
Media aritmética ponderada 62
Propiedades de la media aritmética 63
Cálculo de la media aritmética para datos agrupados 63
La mediana 64
La moda 64
Relación empírica entre la media, la mediana y la moda 64
La media geométrica *G* 65
La media armónica *H* 65
Relación entre las medias aritmética, geométrica y armónica 66
La raíz cuadrada media 66
Cuartiles, deciles y percentiles 66
Software y medidas de tendencia central 67

CAPÍTULO 4 Desviación estándar y otras medidas de dispersión 95
Dispersión o variación 95
Rango 95
Desviación media 95
Rango semiintercuartílico 96
Rango percentil 10-90 96
Desviación estándar 96
Varianza 97
Método abreviado para el cálculo de la desviación estándar 97
Propiedades de la desviación estándar 98
Comprobación de Charlier 99
Corrección de Sheppard para la varianza 100
Relaciones empíricas entre las medidas de dispersión 100
Dispersión absoluta y relativa; coeficiente de variación 100
Variable estandarizada; puntuaciones estándar 101
Software y medidas de dispersión 101

CAPÍTULO 5 Momentos, sesgo y curtosis 123
Momentos 123
Momentos para datos agrupados 123
Relaciones entre momentos 124
Cálculo de momentos con datos agrupados 124
Comprobación de Charlier y corrección de Sheppard 124
Momentos en forma adimensional 124
Sesgo 125
Curtosis 125

Momentos, sesgo y curtosis poblacionales 126
Cálculo del sesgo (o asimetría) y de la curtosis empleando software 126

CAPÍTULO 6 Teoría elemental de la probabilidad 139
Definiciones de probabilidad 139
Probabilidad condicional; eventos independientes y dependientes 140
Eventos mutuamente excluyentes 141
Distribuciones de probabilidad 142
Esperanza matemática 144
Relación entre media y varianza poblacionales y muestrales 144
Análisis combinatorio 145
Combinaciones 146
Aproximación de Stirling para $n!$ 146
Relación entre la probabilidad y la teoría de conjuntos 146
Diagramas de Euler o de Venn y probabilidad 146

CAPÍTULO 7 Las distribuciones binomial, normal y de Poisson 172
La distribución binomial 172
La distribución normal 173
Relación entre las distribuciones binomial y normal 174
La distribución de Poisson 175
Relación entre las distribuciones binomial y de Poisson 176
La distribución multinomial 177
Ajuste de distribuciones teóricas a distribuciones muestrales de frecuencia 177

CAPÍTULO 8 Teoría elemental del muestreo 203
Teoría del muestreo 203
Muestras aleatorias y números aleatorios 203
Muestreo con reposición y sin ella 204
Distribuciones muestrales 204
Distribuciones muestrales de medias 204
Distribuciones muestrales de proporciones 205
Distribuciones muestrales de diferencias y sumas 205
Errores estándar 207
Demostración de la teoría elemental del muestreo empleando software 207

CAPÍTULO 9 Teoría de la estimación estadística 227
Estimación de parámetros 227
Estimaciones insesgadas 227
Estimaciones eficientes 228
Estimaciones puntuales y estimaciones por intervalo; su confiabilidad 228
Estimación de parámetros poblacionales mediante un intervalo de confianza 228
Error probable 230

CAPÍTULO 10 Teoría estadística de la decisión 245
Decisiones estadísticas 245
Hipótesis estadísticas 245
Pruebas de hipótesis y de significancia o reglas de decisión 246
Errores Tipo I y Tipo II 246
Nivel de significancia 246
Pruebas empleando distribuciones normales 246
Pruebas de una y de dos colas 247
Pruebas especiales 248
Curva característica de operación; potencia de una prueba 248
Valor p en pruebas de hipótesis 248
Gráficas de control 249
Pruebas para diferencias muestrales 249
Pruebas empleando distribuciones binomiales 250

CAPÍTULO 11 Teoría de las muestras pequeñas 275
Distribución t de Student 275
Intervalos de confianza 276
Pruebas de hipótesis y de significancia 277
Distribución ji cuadrada 277
Intervalos de confianza para σ 278
Grados de libertad 278
La distribución F 279

CAPÍTULO 12 La prueba ji cuadrada 294
Frecuencias observadas y frecuencias teóricas 294
Definición de χ^2 294
Pruebas de significancia 295
La prueba ji cuadrada de bondad de ajuste 295
Tablas de contingencia 296
Corrección de Yates por continuidad 297
Fórmulas sencillas para calcular χ^2 297
Coeficiente de contingencia 298
Correlación de atributos 298
Propiedad aditiva de χ^2 299

CAPÍTULO 13 Ajuste de curva y método de mínimos cuadrados 316
Relación entre variables 316
Ajuste de curvas 316
Ecuaciones de curvas de aproximación 317
Método de ajuste de curvas a mano 318
La línea recta 318
El método de mínimos cuadrados 319
La recta de mínimos cuadrados 319

Relaciones no lineales 320
La parábola de mínimos cuadrados 320
Regresión 321
Aplicaciones a series de tiempo 321
Problemas en los que intervienen más de dos variables 321

CAPÍTULO 14 Teoría de la correlación 345
Correlación y regresión 345
Correlación lineal 345
Medidas de la correlación 346
Las rectas de regresión de mínimos cuadrados 346
El error estándar de estimación 347
Variación explicada y no explicada 348
Coeficiente de correlación 348
Observaciones acerca del coeficiente de correlación 349
Fórmula producto-momento para el coeficiente de correlación lineal 350
Fórmulas simplificadas para el cálculo 350
Rectas de regresión y el coeficiente de correlación lineal 351
Correlación de series de tiempo 351
Correlación de atributos 351
Teoría muestral de la correlación 351
Teoría muestral de la regresión 352

CAPÍTULO 15 Correlación múltiple y correlación parcial 382
Correlación múltiple 382
Notación empleando subíndice 382
Ecuaciones de regresión y planos de regresión 382
Ecuaciones normales para los planos de regresión de mínimos cuadrados 383
Planos de regresión y coeficientes de correlación 383
Error estándar de estimación 384
Coeficiente de correlación múltiple 384
Cambio de la variable dependiente 384
Generalizaciones a más de tres variables 385
Correlación parcial 385
Relaciones entre coeficientes de correlación múltiple y coeficientes de correlación parcial 386
Regresión múltiple no lineal 386

CAPÍTULO 16 Análisis de varianza 403
Objetivo del análisis de varianza 403
Clasificación en un sentido o experimentos con un factor 403
Variación total, variación dentro de tratamientos y variación entre tratamientos 404
Métodos abreviados para obtener las variaciones 404

Modelo matemático para el análisis de varianza 405
Valores esperados de las variaciones 405
Distribuciones de las variaciones 406
Prueba F para la hipótesis nula de medias iguales 406
Tablas para el análisis de varianza 406
Modificaciones para cantidades desiguales de observaciones 407
Clasificación en dos sentidos o experimentos con dos factores 407
Notación para experimentos con dos factores 408
Variaciones en los experimentos con dos factores 408
Análisis de varianza para experimentos con dos factores 409
Experimentos con dos factores con replicación 410
Diseño experimental 412

CAPÍTULO 17 Pruebas no paramétricas 446
Introducción 446
La prueba de los signos 446
La prueba U de Mann-Whitney 447
La prueba H de Kruskal-Wallis 448
Prueba H corregida para empates 448
Prueba de las rachas para aleatoriedad 449
Otras aplicaciones de la prueba de las rachas 450
Correlación de rangos de Spearman 450

CAPÍTULO 18 Control estadístico de procesos y capacidad de procesos 480
Análisis general de las gráficas de control 480
Gráficas de control de variables y gráficas de control de atributos 481
Gráficas X-barra y gráficas R 481
Pruebas para causas especiales 484
Capacidad de procesos 484
Gráficas P y NP 487
Otras gráficas de control 489

Respuestas a los problemas suplementarios 505

Apéndices 559
I Ordenadas (Y) en z, en la curva normal estándar 561

II Áreas bajo la curva normal estándar, desde 0 hasta z 562

III Valores percentiles (t_p) correspondientes a la distribución t de Student con ν grados de libertad (área sombreada $= p$) 563

IV Valores percentiles (χ^2_p) correspondientes a la distribución Ji cuadrada con ν grados de libertad (área sombreada $= p$) 564

V Valores del percentil 95 correspondientes a la distribución F (ν_1 grados de libertad en el numerador) (ν_2 grados de libertad en el denominador) 565

VI Valores del percentil 99 correspondientes a la distribución F (ν_1 grados de libertad en el numerador) (ν_2 grados de libertad en el denominador) 566

VII Logaritmos comunes con cuatro cifras decimales 567

VIII Valores de $e^{-\lambda}$ 569

IX Números aleatorios 570

Índice **571**

VARIABLES Y GRÁFICAS

1

ESTADÍSTICA

La estadística se ocupa de los métodos científicos que se utilizan para recolectar, organizar, resumir, presentar y analizar datos así como para obtener conclusiones válidas y tomar decisiones razonables con base en este análisis.

El término *estadística* también se usa para denotar los datos o los números que se obtienen de esos datos; por ejemplo, los promedios. Así, se habla de estadísticas de empleo, estadísticas de accidentes, etcétera.

POBLACIÓN Y MUESTRA; ESTADÍSTICA INDUCTIVA (O INFERENCIAL) Y ESTADÍSTICA DESCRIPTIVA

Cuando se recolectan datos sobre las características de un grupo de individuos o de objetos, por ejemplo, estatura y peso de los estudiantes de una universidad o cantidad de pernos defectuosos y no defectuosos producidos en determinado día en una fábrica, suele ser imposible o poco práctico observar todo el grupo, en especial si se trata de un grupo grande. En vez de examinar todo el grupo, al que se le conoce como *población* o *universo*, se examina sólo una pequeña parte del grupo, al que se le llama *muestra*.

Las poblaciones pueden ser *finitas* o *infinitas*. Por ejemplo, la población que consta de todos los pernos producidos determinado día en una fábrica es finita, en tanto que la población que consta de todos los resultados (cara o cruz) que se pueden obtener lanzando una y otra vez una moneda es infinita.

Si la muestra es representativa de la población, el análisis de la muestra permite inferir conclusiones válidas acerca de la población. A la parte de la estadística que se ocupa de las condiciones bajo la cuales tales inferencias son válidas se le llama *estadística inductiva* o *inferencial*. Como estas inferencias no pueden ser absolutamente ciertas, para presentar estas conclusiones se emplea el lenguaje de la probabilidad.

A la parte de la estadística que únicamente trata de describir y analizar un grupo dado, sin sacar ninguna conclusión ni hacer inferencia alguna acerca de un grupo más grande, se le conoce como *estadística descriptiva* o *deductiva*.

Antes de proceder al estudio de la estadística, se analizarán algunos conceptos matemáticos importantes.

VARIABLES: DISCRETAS Y CONTINUAS

Una variable es un símbolo; por ejemplo, X, Y, H, x o B, que puede tomar cualquiera de los valores de determinado conjunto al que se le conoce como *dominio* de la variable. A una variable que sólo puede tomar un valor se le llama *constante*.

Una variable que puede tomar cualquiera de los valores entre dos números dados es una *variable continua*; de lo contrario es una *variable discreta*.

EJEMPLO 1 La cantidad N de hijos que tiene una familia puede tomar los valores 0, 1, 2, 3, . . . , pero no puede tomar valores como 2.5 o 3.842; ésta es una variable discreta.

EJEMPLO 2 La estatura H de una persona que puede ser 62 pulgadas (in), 63.8 in o 65.8341 in, dependiendo de la exactitud con que se mida, es una variable continua.

Los datos descritos mediante una variable discreta son *datos discretos* y los datos descritos mediante una variable continua son *datos continuos*. Un ejemplo de datos discretos es la cantidad de hijos que tiene cada una de 1 000 familias, en tanto que un ejemplo de datos continuos son las estaturas de 100 estudiantes universitarios. En general, una *medición* proporciona datos continuos; en cambio, una *enumeración* o un *conteo* proporciona datos discretos.

Es útil ampliar el concepto de variable a entidades no numéricas; por ejemplo, en el arco iris, color C es una variable que puede tomar los "valores" rojo, anaranjado, amarillo, verde, azul, índigo o violeta. Estas variables se pueden reemplazar por números; por ejemplo, se puede denotar rojo con 1, anaranjado con 2, etcétera.

REDONDEO DE CANTIDADES NUMÉRICAS

El resultado de redondear un número por ejemplo 72.8 a la unidad más cercana es 73 debido a que 72.8 está más cerca de 73 que de 72. De igual manera, 72.8146 redondeado a la centésima más cercana (o a dos lugares decimales) es 72.81, ya que 72.8146 está más cerca de 72.81 que de 72.82.

Sin embargo, para redondear 72.465 a la centésima más cercana, ocurre un dilema debido a que 72.465 se encuentra *precisamente a la mitad* entre 72.46 y 72.47. En estos casos, lo que se acostumbra hacer es redondear al *entero par* antes del 5. Así, 72.465 se redondea a 72.46, 183.575 se redondea a 183.58 y 116 500 000, redondeado al millón más cercano, es 116 000 000. Hacer esto es especialmente útil cuando se realiza una gran cantidad de operaciones para minimizar, así, el *error de redondeo acumulado* (ver problema 1.4).

NOTACIÓN CIENTÍFICA

Al escribir números, en especial aquellos en los que hay muchos ceros antes o después del punto decimal, es conveniente usar la notación científica empleando potencias de 10.

EJEMPLO 3 $10^1 = 10$, $10^2 = 10 \times 10 = 100$, $10^5 = 10 \times 10 \times 10 \times 10 \times 10 = 100\,000$ y $10^8 = 100\,000\,000$.

EJEMPLO 4 $10^0 = 1$, $10^{-1} = .1$ o 0.1; $10^{-2} = .01$ o 0.01; y $10^{-5} = .00001$ o 0.00001.

EJEMPLO 5 $864\,000\,000 = 8.64 \times 10^8$ y $0.00003416 = 3.416 \times 10^{-5}$.

Obsérvese que el efecto de multiplicar un número, por ejemplo, por 10^8, es recorrer el punto decimal del número ocho lugares *a la derecha*. El efecto de multiplicar un número por 10^{-6} es recorrer el punto decimal del número seis lugares *a la izquierda*.

Con frecuencia, para hacer énfasis en que no se ha omitido un número distinto de cero antes del punto decimal, se escribe 0.1253 en lugar de .1253. Sin embargo, en casos en los que no pueda haber lugar a confusión, como en tablas, el cero antes del punto decimal puede omitirse.

Para indicar la multiplicación de dos o más números se acostumbra usar paréntesis o puntos. Así $(5)(3) = 5 \cdot 3 = 5 \times 3 = 15$, y $(10)(10)(10) = 10 \cdot 10 \cdot 10 = 10 \times 10 \times 10 = 1\,000$. Cuando se utilizan letras para representar números suelen omitirse los paréntesis y los puntos; por ejemplo, $ab = (a)(b) = a \cdot b = a \times b$.

La notación científica es útil al hacer cálculos, en especial para localizar el punto decimal. Entonces se hace uso de las reglas siguientes:

$$(10^p)(10^q) = 10^{p+q} \qquad \frac{10^p}{10^q} = 10^{p-q}$$

donde p y q son números cualesquiera.

En 10^p, p es el *exponente* y 10 es la *base*.

EJEMPLO 6 $(10^3)(10^2) = 1\,000 \times 100 = 100\,000 = 10^5$ es decir, 10^{3+2}

$$\frac{10^6}{10^4} = \frac{1\,000\,000}{10\,000} = 100 = 10^2 \quad \text{es decir, } 10^{6-4}$$

EJEMPLO 7 $(4\,000\,000)(0.0000000002) = (4 \times 10^6)(2 \times 10^{-10}) = (4)(2)(10^6)(10^{-10}) = 8 \times 10^{6-10}$

$$= 8 \times 10^{-4} = 0.0008$$

EJEMPLO 8 $$\frac{(0.006)(80\,000)}{0.04} = \frac{(6 \times 10^{-3})(8 \times 10^4)}{4 \times 10^{-2}} = \frac{48 \times 10^1}{4 \times 10^{-2}} = \left(\frac{48}{4}\right) \times 10^{1-(-2)}$$

$$= 12 \times 10^3 = 12\,000$$

CIFRAS SIGNIFICATIVAS

Si se anota la estatura de una persona como 65.4 in, esto significa que la estatura verdadera estará entre 65.35 y 65.45 in. Los dígitos exactos, fuera de los ceros necesarios para localizar el punto decimal, son los *dígitos significativos* o *cifras significativas* del número.

EJEMPLO 9 65.4 tiene tres cifras significativas.

EJEMPLO 10 4.5300 tiene cinco cifras significativas.

EJEMPLO 11 $.0018 = 0.0018 = 1.8 \times 10^{-3}$ tiene dos cifras significativas.

EJEMPLO 12 $.001800 = 0.001800 = 1.800 \times 10^{-3}$ tiene cuatro cifras significativas.

Los números obtenidos de enumeraciones (o conteos), a diferencia de los obtenidos de mediciones, por supuesto son exactos y por lo tanto tienen un número ilimitado de cifras significativas. Sin embargo, en algunos de estos casos puede ser difícil decidir, sin más información, cuáles cifras son significativas. Por ejemplo, el número 186 000 000 puede tener 3, 4, . . . , 9 cifras significativas. Si se sabe que tiene cinco cifras significativas puede ser más adecuado escribirlo como 186.00 millones o como 1.8600×10^8.

CÁLCULOS

Al realizar cálculos en los que intervienen multiplicaciones, divisiones o raíces de números, el resultado final no puede tener más cifras significativas que el número con menos cifras significativas (ver problema 1.9).

EJEMPLO 13 $73.24 \times 4.53 = (73.24)(4.52) = 331$

EJEMPLO 14 $1.648/0.023 = 72$

EJEMPLO 15 $\sqrt{38.7} = 6.22$

EJEMPLO 16 $(8.416)(50) = 420.8$ (si 50 es exacto)

Cuando se suman o restan números, el resultado final no puede tener más cifras significativas después del punto decimal que los números con menos cifras significativas después del punto decimal (ver problema 1.10).

EJEMPLO 17 $3.16 + 2.7 = 5.9$

EJEMPLO 18 $83.42 - 72 = 11$

EJEMPLO 19 $47.816 - 25 = 22.816$ (si 25 es exacto)

La regla anterior para la suma y la resta puede extenderse (ver problema 1.11).

FUNCIONES

Si a cada valor que puede tomar la variable X le corresponde un valor de una variable Y, se dice que Y es *función* de X y se escribe $Y = F(X)$ (se lee "Y es igual a F de X") para indicar esta dependencia funcional. En lugar de F también pueden usarse otras letras (G, ϕ, etcétera).

La variable X es la *variable independiente* y la variable Y es la *variable dependiente*.

Si a cada valor de X le corresponde únicamente un valor de Y, se dice que Y es una *función univaluada* de X; de lo contrario, se dice que es una *función multivaluada* de X.

EJEMPLO 20 La población P de Estados Unidos es función del tiempo t, lo que se escribe $P = F(t)$.

EJEMPLO 21 El estiramiento S de un resorte vertical es función del peso W que hay en el extremo del resorte, es decir, $S = G(W)$.

La dependencia (o correspondencia) funcional entre variables puede describirse mediante una tabla. Pero también puede indicarse mediante una ecuación que relaciona las variables, por ejemplo, $Y = 2X - 3$, a partir de la cual puede determinarse el valor de Y que corresponde a los diversos valores de X.

Si $Y = F(X)$, $F(3)$ denota "el valor de Y cuando $X = 3$", $F(10)$ denota "el valor de Y cuando $X = 10$", etc. Así, si $Y = F(X) = X^2$, entonces, $F(3) = 3^2 = 9$ es el valor de Y cuando $X = 3$.

El concepto de función puede ampliarse a dos o más variables (ver problema 1.17).

COORDENADAS RECTANGULARES

En la figura 1-1 se muestra un diagrama de dispersión de EXCEL con cuatro puntos. Este *diagrama de dispersión* está formado por dos rectas mutuamente perpendiculares llamadas *ejes* X y Y. El eje X es horizontal y el eje Y es vertical. Estos dos ejes se cortan en un punto llamado *origen*. Estas dos rectas dividen al *plano XY* en cuatro regiones que se denotan I, II, II y IV, a las que se les conoce como primero, segundo, tercero y cuarto *cuadrantes*. En la figura 1-1 se muestran cuatro puntos. El punto (2, 3) está en el primer cuadrante y se grafica avanzando, desde el origen, 2 unidades a la derecha sobre el eje X y desde ahí, 3 unidades hacia arriba. El punto $(-2.3, 4.5)$ está en el segundo cuadrante y se grafica avanzando, desde el origen, 2.3 unidades a la izquierda sobre el eje X y desde ahí, 4.5 unidades hacia arriba. El punto $(-4, -3)$ está en el tercer cuadrante y se grafica avanzando, desde el origen, 4 unidades a la izquierda sobre el eje X, y desde ahí 3 unidades hacia abajo. El punto $(3.5, -4)$ está en el cuarto cuadrante y se grafica avanzando 3.5 unidades a la derecha sobre el eje X, y desde ahí 4 unidades hacia abajo. El primer número de cada uno de estos pares es la *abscisa* del punto y el segundo número es la *ordenada* del punto. La abscisa y la ordenada, juntas, son las *coordenadas* del punto.

Las ideas anteriores pueden ampliarse construyendo un eje Z a través del origen y perpendicular al plano XY. En este caso las coordenadas de cada punto se denotan (X, Y, Z).

GRÁFICAS

Una *gráfica* es una representación visual de la relación entre las variables. En estadística, dependiendo de la naturaleza de los datos y del propósito que se persiga, se emplean distintos tipos de gráficas: *gráficas de barras*, *de pastel*, *pictogramas*, etc. A las gráficas también se les suele llamar *cartas* o *diagramas*. Así, se habla de cartas de barras, diagramas de pastel, etc. (ver los problemas 1.23, 1.24, 1.25, 1.26 y 1.27).

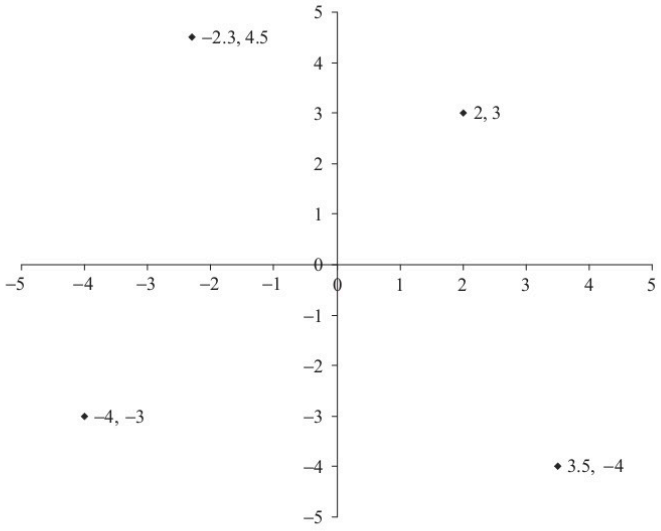


Figura 1-1 EXCEL, gráfica de puntos en los cuatro cuadrantes.

ECUACIONES

Las ecuaciones son expresiones de la forma $A = B$, donde A es el *miembro* (o *lado*) *izquierdo* de la ecuación y B es el *miembro* (o *lado*) *derecho*. Si se aplican las mismas operaciones a ambos lados de una ecuación se obtienen *ecuaciones equivalentes*. Así, si a ambos miembros de una ecuación se suma o resta un mismo número se obtiene una ecuación equivalente; también, si ambos lados se multiplican por un mismo número o se dividen entre un mismo número, con excepción de la *división entre cero que no es válida*, se obtiene una ecuación equivalente.

EJEMPLO 22 Dada la ecuación $2X + 3 = 9$, se resta 3 a ambos miembros: $2X + 3 - 3 = 9 - 3$ o $2X = 6$. Se dividen ambos miembros entre 2: $2X/2 = 6/2$ o $X = 3$. Este valor de X es una solución de la ecuación dada, como se puede ver sustituyendo X por 3, con lo que se obtiene $2(3) + 3 = 9$, o $9 = 9$, que es una *identidad*. Al proceso de obtener las soluciones de una ecuación se le conoce como *resolver* la ecuación.

Las ideas anteriores pueden extenderse a hallar soluciones de dos ecuaciones en dos incógnitas, de tres ecuaciones en tres incógnitas, etc. A tales ecuaciones se les conoce como *ecuaciones simultáneas* (ver problema 1.30).

DESIGUALDADES

Los símbolos $<$ y $>$ significan "menor que" y "mayor que", respectivamente. Los símbolos $\leq$ y $\geq$ significan "menor o igual a" y "mayor o igual a", respectivamente. Todos estos símbolos se conocen como *signos de desigualdad*.

EJEMPLO 23 $3 < 5$ se lee "3 es menor que 5".

EJEMPLO 24 $5 > 3$ se lee "5 es mayor que 3".

EJEMPLO 25 $X < 8$ se lee "X es menor que 8".

EJEMPLO 26 $X \geq 10$ se lee "X es mayor o igual a 10".

EJEMPLO 27 $4 < Y \leq 6$ se lee "4 es menor que Y y Y es menor o igual a 6" o "Y está entre 4 y 6, excluyendo al 4 e incluyendo al 6" o "Y es mayor que 4 y menor o igual a 6".

A las relaciones en las que intervienen signos de desigualdad se les llana *desigualdades*. Así como se habla de miembros de una ecuación, también se habla de *miembros de una desigualdad*. Por lo tanto, en la desigualdad $4 < Y \leq 6$, los miembros son 4, Y y 6.

Una desigualdad válida sigue siendo válida si:

1. A cada miembro de la desigualdad se le suma o se le resta un mismo número.

EJEMPLO 28 Como $15 > 12$, $15 + 3 > 12 + 3$ (es decir, $18 > 15$) y $15 - 3 > 12 - 3$ (es decir, $12 > 9$).

2. Cada miembro de la desigualdad se multiplica por un mismo número *positivo* o se divide entre un mismo número positivo.

EJEMPLO 29 Como $15 > 12$, $(15)(3) > (12)(3)$ (es decir, $45 > 36$) y $15/3 > 12/3$ (es decir, $5 > 4$).

3. Cada miembro se multiplica o se divide por un mismo número *negativo*, lo que indica que los símbolos de la desigualdad son invertidos.

EJEMPLO 30 Como $15 > 12$, $(15)(-3) < (12)(-3)$ (es decir, $-45 < -36$) y $15/(-3) < 12/(-3)$ (es decir, $-5 < -4$).

LOGARITMOS

Si $x > 0$, $b > 0$ y $b \neq 1$, $y = \log_b x$ si y sólo si $\log b^y = x$. Un logaritmo es un exponente. Es la potencia a la que hay que elevar la base b para obtener el número del que se busca el logaritmo. Las dos bases más utilizadas son el 10 y la e, que es igual a 2.71828182... A los logaritmos base 10 se les llama *logaritmos comunes* y se escriben $\log_{10} x$ o simplemente $\log(x)$. A los logaritmos base e se les llama *logaritmos naturales* y se escriben $\ln(x)$.

EJEMPLO 31 Encuentre los siguientes logaritmos y después encuéntrelos usando EXCEL: $\log_2 8$, $\log_5 25$ y $\log_{10} 1\,000$. La potencia a la que hay que elevar al 2 para obtener 8 es tres, así $\log_2 8 = 3$. La potencia a la que hay que elevar al 5 para obtener 25 es dos, así $\log_5 25 = 2$. La potencia a la que hay que elevar al 10 para obtener 1 000 es tres, así $\log_{10} 1\,000 = 3$. EXCEL tiene tres funciones para calcular logaritmos. La función LN calcula logaritmos naturales, la función LOG10 calcula logaritmos comunes y la función LOG(x,b) calcula el logaritmo de x base b. =LOG(8,2) da 3, =LOG(25,5) da 2, =LOG10(1 000) da 3.

EJEMPLO 32 Calcule los logaritmos naturales de los números del 1 al 5 usando EXCEL. Los números 1 a 5 se ingresan en las celdas B1:F1 y en la celda B2 se ingresa la expresión =LN(B1), se hace clic y se arrastra desde B2 hasta F2. EXCEL proporciona el siguiente resultado.

X	1	2	3	4	5
LN(x)	0	0.693147	1.098612	1.386294	1.609438

EJEMPLO 33 Muestre que las respuestas del ejemplo 32 son correctas mostrando que $e^{\ln(x)}$ da el valor x. Los logaritmos se ingresan en B1:F1 y la expresión $e^{\ln(x)}$, que está representada por =EXP(B1) se ingresa en B2, se hace clic y se arrastra de B2 a F2. EXCEL da los resultados siguientes. Los números en D2 y E2 difieren de 3 y 4 debido a error de redondeo.

LN(x)	0	0.693147	1.098612	1.386294	1.609438
x=EXP(LN(x))	1	2	2.999999	3.999999	5

El ejemplo 33 ilustra que si se tiene el logaritmo de un número ($\log_b(x)$) se puede volver a obtener el número x usando la relación $b^{\log_b(x)} = x$.

EJEMPLO 34 El número e puede definirse como un límite. La cantidad $(1 + (1/x))^x$ se va acercando a e a medida que x va creciendo. Obsérvense las evaluaciones de EXCEL de $(1 + (1/x))^x$ para $x = 1, 10, 100, 1\,000, 10\,000, 100\,000$ y $1\,000\,000$.

x	1	10	100	1 000	10 000	100 000	1 000 000
(1+1/x)^x	2	2.593742	2.704814	2.716924	2.718146	2.718268	2.71828

Los números 1, 10, 100, 1 000, 10 000, 100 000 y 1 000 000 se ingresan en B1:H1 y la expresión = (1 + 1/B1)^B1 se ingresa en B2, se hace clic y se arrastra de B2 a H2. Esto se expresa matemáticamente mediante la expresión $\lim_{x\to\infty}(1 + (1/x))^x = e$.

EJEMPLO 35 El saldo de una cuenta que gana interés compuesto n veces por año está dado por $A(t) = P(1 + (r/n))^{nt}$ donde P es el capital, r es la tasa de interés, t es el tiempo en años y n es el número de periodos compuestos por año. El saldo de una cuenta que gana interés continuo está dado por $A(t) = Pe^{rt}$. Para comparar el crecimiento de $1 000 a interés continuo con el de $1 000 a interés compuesto trimestralmente, después de 1, 2, 3, 4 y 5 años, ambos a una tasa de interés de 5%, se usa EXCEL. Los resultados son:

Años	1	2	3	4	5
Trimestralmente	1 050.95	1 104.49	1 160.75	1 219.89	1 282.04
Continuamente	1 051.27	1 105.17	1 161.83	1 221.4	1 284.03

Se ingresan los tiempos 1, 2, 3, 4 y 5 en B1:F1; en B2 se ingresa la expresión de EXCEL =1 000*(1.0125)^(4*B1), se hace clic y se arrastra desde B2 hasta F2. En B3 se ingresa la expresión =1 000*EXP(0.05*B1), se hace clic y se arrastra desde B3 hasta F3. El interés continuo compuesto da resultados ligeramente mejores.

PROPIEDADES DE LOS LOGARITMOS

Las propiedades más importantes de los logaritmos son las siguientes:

1. $\log_b MN = \log_b M + \log_b N$
2. $\log_b M/N = \log_b M - \log_b N$
3. $\log_b M^P = p \log_b M$

EJEMPLO 36 Escriba $\log_b(xy^4/z^3)$ como suma o diferencia de logaritmos de x, y y z.

$$\log_b \frac{xy^4}{z^3} = \log_b xy^4 - \log_b z^3 \quad \text{propiedad 2}$$

$$\log_b \frac{xy^4}{z^3} = \log_b x + \log_b y^4 - \log_b z^3 \quad \text{propiedad 1}$$

$$\log_b \frac{xy^4}{z^3} = \log_b x + 4\log_b y - 3\log_b z \quad \text{propiedad 3}$$

ECUACIONES LOGARÍTMICAS

Para resolver ecuaciones logarítmicas:

1. Todos los logaritmos se aíslan en un lado de la ecuación.
2. Las sumas o diferencias de logaritmos se expresan como un solo logaritmo.
3. La ecuación obtenida en el paso 2 se expresa en forma exponencial.
4. Se resuelve la ecuación obtenida en el paso 3.
5. Se verifican las soluciones.

EJEMPLO 37 Solucione la siguiente ecuación logarítmica: $\log_4(x+5)=3$. Primero, se expresa esta ecuación en forma exponencial como $x+5=4^3=64$. A continuación se despeja x como sigue, $x=64-5=59$. Por último se verifica la solución. $\log_4(59+5)=\log_4(64)=3$ ya que $4^3=64$.

EJEMPLO 38 Resuelva la ecuación logarítmica siguiente: $\log(6y-7)+\log y=\log(5)$. La suma de logaritmos se reemplaza como el logaritmo del producto, $\log(6y-7)y=\log(5)$. Se igualan $(6y-7)y$ y 5. El resultado es $6y^2-7y=5$ o $6y^2-7y-5=0$. Se factoriza esta ecuación cuadrática como $(3y-5)(2y+1)=0$. Las soluciones son $y=5/3$ y $y=-1/2$. El $-1/2$ se descarta como solución, ya que los logaritmos de números negativos no están definidos. $y=5/3$ demuestra ser una solución cuando se sustituye en la ecuación original. Por lo tanto, la única solución es $y=5/3$.

EJEMPLO 39 Resuelva la ecuación logarítmica siguiente:

$$\ln(5x)-\ln(4x+2)=4$$

La diferencia de logaritmos se convierte en el logaritmo del cociente, $\ln(5x/(4x+2))=4$. Aplicando la definición de logaritmo: $5x/(4x+2)=e^4=54.59815$. Despejando x de la ecuación $5x=218.39260x+109.19630$ se obtiene $x=-0.5117$. Sin embargo, esta respuesta no satisface la ecuación $\ln(5x)-\ln(4x+2)=4$, ya que la función log no está definida para números negativos. La ecuación $\ln(5x)-\ln(4x+2)=4$ no tiene solución.

PROBLEMAS RESUELTOS

VARIABLES

1.1 En cada uno de los casos siguientes indíquese si se trata de datos continuos o de datos discretos:

a) Cantidad de acciones que se venden diariamente en la bolsa de valores.
b) Temperatura registrada cada media hora en un observatorio.
c) Vida media de los cinescopios producidos por una empresa.
d) Ingreso anual de los profesores universitarios.
e) Longitud de 100 pernos producidos en una fábrica

SOLUCIÓN

a) Discreta; *b*) continua; *c*) continua; *d*) discreta; *e*) continua.

1.2 Dar el dominio de cada una de las variables siguiente e indicar si es una variable continua o discreta.

a) Cantidad G de galones (gal) de agua en una lavadora.
b) Cantidad B de libros en un anaquel.
c) Suma S de la cantidad de puntos que se obtienen al lanzar un par de dados.
d) Diámetro D de una esfera.
e) País C en Europa.

SOLUCIÓN

a) *Dominio:* Cualquier valor desde 0 gal hasta la capacidad de la máquina. *Variable:* continua.
b) *Dominio:* 0, 1, 2, 3, . . . hasta la mayor cantidad de libros que se quepan en el anaquel. *Variable:* discreta.
c) *Dominio:* Con un solo dado se pueden obtener 1, 2, 3, 4, 5 o 6 puntos. Por lo tanto, la suma de puntos en un par de dados puede ser 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 y 12, los cuales constituyen el dominio de S. *Variable:* discreta.
d) *Dominio:* Si se considera un punto como una esfera de diámetro cero, el dominio de D son todos los valores desde cero en adelante. *Variable:* continua.
e) *Dominio:* Inglaterra, Francia, Alemania, etc., que pueden representarse por medio de los números 1, 2, 3, etc. *Variable:* discreta.

REDONDEO DE CANTIDADES NUMÉRICAS

1.3 Redondear cada uno de los números siguientes como se indica:

a)	48.6	a la unidad más cercana	*f*)	143.95	a la décima más cercana
b)	136.5	a la unidad más cercana	*g*)	368	a la centena más cercana
c)	2.484	a la centésima más cercana	*h*)	24 448	al millar más cercano
d)	0.0435	a la milésima más cercana	*i*)	5.56500	a la centésima más cercana
e)	4.50001	a la unidad más cercana	*j*)	5.56501	a la centésima más cercana

SOLUCIÓN

a) 49; *b*) 136; *c*) 2.48; *d*) 0.044; *e*) 5; *f*) 144.0; *g*) 400; *h*) 24 000; *i*) 5.56; *j*) 5.57

1.4 Sumar los números 4.35, 8.65, 2.95, 12.45, 6.65, 7.55 y 9.75: *a*) directamente, *b*) redondeando a la décima más cercana de acuerdo con la convención del "entero par" y *c*) redondeando de manera que se incremente el dígito antes del 5.

SOLUCIÓN

a)		*b*)		*c*)	
	4.35		4.4		4.4
	8.65		8.6		8.7
	2.95		3.0		3.0
	12.45		12.4		12.5
	6.65		6.6		6.7
	7.55		7.6		7.6
	9.75		9.8		9.8
Total	52.35	Total	52.4	Total	52.7

Obsérvese que el procedimiento *b*) es mejor que el procedimiento *c*) debido a que en el procedimiento *b*) se minimiza la *acumulación de errores de redondeo*.

NOTACIÓN CIENTÍFICA Y CIFRAS SIGNIFICATIVAS

1.5 Expresar cada uno de los números siguiente sin utilizar potencias de 10.

a) 4.823×10^{7} *c*) 3.8×10^{-7} *e*) 300×10^{8}
b) 8.4×10^{-6} *d*) 1.86×10^{5} *f*) $70\,000 \times 10^{-10}$

SOLUCIÓN

a) Se recorre el punto decimal siete lugares a la derecha y se obtiene 48 230 000; *b*) se recorre el punto decimal seis lugares a la izquierda y se obtiene 0.0000084; *c*) 0.000380; *d*) 186 000; *e*) 30 000 000 000; *f*) 0.0000070000.

1.6 En cada inciso diga cuántas cifras significativas hay, entendiéndose que los números se han dado exactamente.

a) 149.8 in *d*) 0.00280 m *g*) 9 casas
b) 149.80 in *e*) 1.00280 m *h*) 4.0×10^{3} libras (lb)
c) 0.0028 metros (m) *f*) 9 gramos (g) *i*) 7.58400×10^{-5} dinas

SOLUCIÓN

a) Cuatro; *b*) cinco; *c*) dos; *d*) tres; *e*) seis; *f*) una; *g*) ilimitadas; *h*) dos; *i*) seis.

1.7 ¿Cuál es el error máximo en cada una de las mediciones siguientes, entendiéndose que se han registrado exactamente?

a) 73.854 in *b*) 0.09800 pies cúbicos (ft^3) *c*) 3.867×10^{8} kilómetros (km)

SOLUCIÓN

a) Esta medida puede variar desde 73.8535 hasta 73.8545 in; por lo tanto, el error máximo es 0.0005 in. Hay cinco cifras significativas.

b) La cantidad de pies cúbicos puede variar desde 0.097995 hasta 0.098005 pies cúbicos; por lo tanto, el error máximo es 0.0005 ft^3. Hay cuatro cifras significativas.

c) El verdadero número de kilómetros es mayor que 3.8665×10^{8}, pero menor que 3.8675×10^{8}; por lo tanto, el error máximo es 0.0005×10^{8}, o 50 000 km. Hay cuatro cifras significativas.

1.8 Escribir cada número empleando la notación científica. A menos que se indique otra cosa, supóngase que todas las cifras son significativas.

a) 24 380 000 (cuatro cifras significativas)
b) 0.000009851
c) 7 300 000 000 (cinco cifras significativas)
d) 0.00018400

SOLUCIÓN

a) 2.438×10^7; *b*) 9.851×10^{-6}; *c*) 7.30000×10^9; *d*) 1.8400×10^{-4}

CÁLCULOS

1.9 Mostrar que el producto de los números 5.74 y 3.8, entendiéndose que tienen tres y dos cifras significativas, respectivamente, no puede ser exacto a más de dos cifras significativas.

SOLUCIÓN

Primer método

$5.74 \times 3.8 = 21.812$, pero en este producto no todas las cifras son significativas. Para determinar cuántas cifras son significativas, obsérvese que 5.74 representa algún número entre 5.735 y 5.745, y 3.8 representa algún número entre 3.75 y 3.85. Por lo tanto, el menor valor que puede tener este producto es $5.735 \times 3.75 = 21.50625$ y el mayor valor que puede tener es $5.745 \times 3.85 = 22.11825$.

Dado que este intervalo de valores es 21.50625 a 22.11825, es claro que sólo los dos primeros dígitos del producto son significativos y el resultado se escribe como 22. Nótese que el número 22 se determina para cualquier número entre 21.5 y 22.5.

Segundo método

Imprimiendo en cursivas las cifras dudosas, este producto se puede calcular como sigue:

$$\begin{array}{r} 5.7\,4 \\ \times\quad 3\,8 \\ \hline 4\,5\,9\,2 \\ 1\,7\,2\,2 \\ \hline 2\,1.8\,1\,2 \end{array}$$

En el resultado no se debe conservar más de una cifra dudosa, por lo que el resultado es 22 a dos cifras significativas. Obsérvese que no es necesario trabajar con más cifras significativas que las presentes en el factor menos exacto; por lo tanto, si 5.74 se redondea a 5.7, el producto será $5.7 \times 3.8 = 21.66 = 22$, a dos cifras significativas, lo cual coincide con el resultado obtenido antes.

Cuando los cálculos se hacen sin calculadora, se puede ahorrar trabajo si no se conserva más de una o dos cifras más de las que tiene el factor menos exacto y se redondea el resultado al número adecuado de cifras significativas. Cuando se usa una computadora, que puede dar muchos dígitos, hay que tener cuidado de no creer que todos los dígitos son significativos.

1.10 Sume los números 4.19355, 15.28, 5.9561, 12.3 y 8.472, entendiéndose que todas las cifras son significativas.

SOLUCIÓN

En el cálculo *a*), que se presenta en la página siguiente, las cifras dudosas están en cursivas. El resultado final con no más de una cifra dudosa es 46.2

$$
\begin{array}{rr}
a) & 4.19355 \\
 & 15.28 \\
 & 5.9561 \\
 & 12.3 \\
 & 8.472 \\
\hline
 & 46.20165
\end{array}
\qquad
\begin{array}{rr}
b) & 4.19 \\
 & 15.28 \\
 & 5.96 \\
 & 12.3 \\
 & 8.47 \\
\hline
 & 46.20
\end{array}
$$

Se puede ahorrar un poco de trabajo si se hacen los cálculos como en el inciso *b*), donde únicamente se ha conservado un lugar decimal más de los que tiene el número menos exacto. El resultado final se redondea a 46.2, que coincide con el resultado en el inciso *a*).

1.11 Calcular 475 000 000 + 12 684 000 − 1 372 410 si estos números tienen tres, cinco y siete cifras significativas, respectivamente.

SOLUCIÓN

En el cálculo *a*) que se muestra abajo, se conservan todas las cifras y se redondea el resultado final. En el cálculo se usa un método similar al del problema 1.10 *b*). En ambos casos las cifras dudosas aparecen en cursivas.

$$
\begin{array}{rr}
a) & 475\,000\,000 \\
 & +\ 12\,684\,000 \\
\hline
 & 487\,684\,000
\end{array}
\qquad
\begin{array}{r}
487\,684\,000 \\
-\ 1\,372\,410 \\
\hline
486\,311\,590
\end{array}
\qquad
\begin{array}{rr}
b) & 475\,000\,000 \\
 & +\ 12\,700\,000 \\
\hline
 & 487\,700\,000
\end{array}
\qquad
\begin{array}{r}
487\,700\,000 \\
-\ 1\,400\,000 \\
\hline
486\,300\,000
\end{array}
$$

El resultado final se redondea a 486 000 000; o mejor aún, para indicar que hay tres cifras significativas, se escribe 486 millones o 4.86×10^8.

1.12. Realizar las operaciones siguientes

a) 48.0×943

b) $8.35/98$

c) $(28)(4\,193)(182)$

d) $\dfrac{(526.7)(0.001280)}{0.000034921}$

e) $\dfrac{(1.47562 - 1.47322)(4\,895.36)}{0.000159180}$

f) Si los denominadores 5 y 6 son exactos, $\dfrac{(4.38)^2}{5} + \dfrac{(5.482)^2}{6}$

g) $3.1416\sqrt{71.35}$

h) $\sqrt{128.5 - 89.24}$

SOLUCIÓN

a) $48.0 \times 943 = (48.0)(943) = 45\,300$

b) $8.35/98 = 0.085$

c) $(28)(4\,193)(182) = (2.8 \times 10^1)(4.193 \times 10^3)(1.82 \times 10^2)$
$= (2.8)(4.193)(1.82) \times 10^{1+3+2} = 21 \times 10^6 = 2.1 = 10^7$

Lo que también puede escribirse como 21 millones, para indicar que hay dos cifras significativas.

d) $\dfrac{(526.7)(0.001280)}{0.000034921} = \dfrac{(5.267 \times 10^2)(1.280 \times 10^{-3})}{3.4921 \times 10^{-5}} = \dfrac{(5.267)(1.280)}{3.4921} \times \dfrac{(10^2)(10^{-3})}{10^{-5}}$

$$= 1.931 \times \frac{10^{2-3}}{10^{-5}} = 1.931 \times \frac{10^{-1}}{10^{-5}}$$

$$= 1.931 \times 10^{-1+5} = 1.931 \times 10^4$$

Lo que también se puede escribir como 19.31 miles, para indicar que hay cuatro cifras significativas.

e) $$\frac{(1.47562 - 1.47322)(4\,895.36)}{0.000159180} = \frac{(0.00240)(4\,895.36)}{0.000159180} = \frac{(2.40 \times 10^{-3})(4.89536 \times 10^{3})}{1.59180 \times 10^{-4}}$$

$$= \frac{(2.40)(4.89536)}{1.59180} \times \frac{(10^{-3})(10^{3})}{10^{-4}} = 7.38 \times \frac{10^{0}}{10^{-4}} = 7.38 \times 10^{4}$$

Lo que también se puede escribir como 73.8 miles para indicar que hay tres cifras significativas. Obsérvese que aunque originalmente en todos los números había seis cifras significativas, al sustraer 1.47322 de 1.47562 algunas de estas cifras significativas se perdieron.

f) Si los denominadores 5 y 6 son exactos, $\frac{(4.38)^2}{5} = \frac{(5.482)^2}{6} = 3.84 + 5.009 = 8.85$

g) $3.1416\sqrt{71.35} = (3.1416)(8.447) = 26.54$

h) $\sqrt{128.5 - 89.24} = \sqrt{39.3} = 6.27$

1.13 Evaluar cada una de las expresiones siguientes, con $X = 3$, $Y = -5$, $A = 4$ y $B = -7$, donde todos los números se supone que son exactos:

a) $2X - 3Y$

b) $4Y - 8X + 28$

c) $\dfrac{AX + BY}{BX - AY}$

d) $X^2 - 3XY - 2Y^2$

e) $2(X + 3Y) - 4(3X - 2Y)$

f) $\dfrac{X^2 - Y^2}{A^2 - B^2 + 1}$

g) $\sqrt{2X^2 - Y^2 - 3A^2 + 4B^2 + 3}$

h) $\sqrt{\dfrac{6A^2}{X} + \dfrac{2B^2}{Y}}$

SOLUCIÓN

a) $2X - 3Y = 2(3) - 3(-5) = 6 + 15 = 21$

b) $4Y - 8X + 28 = 4(-5) - 8(3) + 28 = -20 - 24 + 28 = -16$

c) $\dfrac{AX + BY}{BX - AY} = \dfrac{(4)(3) + (-7)(-5)}{(-7)(3) - (4)(-5)} = \dfrac{12 + 35}{-21 + 20} = \dfrac{47}{-1} = -47$

d) $X^2 - 3XY - 2Y^2 = (3)^2 - 3(3)(-5) - 2(-5)^2 = 9 + 45 - 50 = 4$

e) $2(X + 3Y) - 4(3X - 2Y) = 2[(3) + 3(-5)] - 4[3(3) - 2(-5)]$

$$= 2(3 - 15) - 4(9 + 10) = 2(-12) - 4(19) = -24 - 76 = -100$$

Otro método

$$2(X + 3Y) - 4(3X - 2Y) = 2X + 6Y - 12X + 8Y = -10X + 14Y = -10(3) = 14(-5)$$

$$= -30 - 70 = -100$$

f) $\dfrac{X^2 - Y^2}{A^2 - B^2 + 1} = \dfrac{(3)^2 - (-5)^2}{(4)^2 - (-7)^2 + 1} = \dfrac{9 - 25}{16 - 49 + 1} = \dfrac{-16}{-32} + \dfrac{1}{2} = 0.5$

g) $\sqrt{2X^2 - Y^2 - 3A^2 + 4B^2 + 3} = \sqrt{2(3)^2 - (-5)^2 - 3(4)^2 + 4(-7)^2 + 3}$

$$= \sqrt{18 - 25 - 48 + 196 + 3} = \sqrt{144} = 12$$

h) $\sqrt{\dfrac{6A^2}{X} + \dfrac{2B^2}{Y}} = \sqrt{\dfrac{6(4)^2}{3} + \dfrac{2(-7)^2}{-5}} = \sqrt{\dfrac{96}{3} + \dfrac{98}{-5}} = \sqrt{12.4} = 3.52$ aproximadamente

FUNCIONES Y GRÁFICAS

1.14 En la tabla 1.1 se presentan las cantidades de bushels (bu) de trigo y de maíz producidas en una granja en los años 2002, 2003, 2004, 2005 y 2006. De acuerdo con esta tabla, determinar el año o los años en los que: *a*) se produjeron menos bushels de trigo, *b*) se produjo la mayor cantidad de bushels de maíz, *c*) hubo la mayor dis-

Tabla 1.1 Producción de trigo y maíz desde 2002 hasta 2006

Año	Bushels de trigo	Bushels de maíz
2002	205	80
2003	215	105
2004	190	110
2005	205	115
2006	225	120

minución en la producción de trigo, *d*) se produjo una misma cantidad de trigo, *e*) la suma de las producción de trigo y maíz fue máxima.

SOLUCIÓN

a) 2004; *b*) 2006; *c*) 2004; *d*) 2002 y 2005; *e*) 2006

1.15 Sean W y C, respectivamente, las cantidades de bushels de trigo y maíz producidas en el año t en la granja del problema 1.14. Es claro que W y C son funciones de t; esto se indica como $W = F(t)$ y $C = G(t)$.

a) Encontrar W para $t = 2004$.
b) Encontrar C para $t = 2002$.
c) Encontrar t para $W = 205$.
d) Encontrar $F(2005)$.
e) Encontrar $G(2005)$.
f) Encontrar C para $W = 190$.
g) ¿Cuál es el dominio de la variable t?
h) ¿Es W una función univaluada de t?
i) ¿Es t función de W?
j) ¿Es C función de W?
k) ¿Cuál es una variable independiente, t o W?

SOLUCIÓN

a) 190
b) 80
c) 2002 y 2005
d) 205
e) 115
f) 110
g) Todos los años, desde el 2002 hasta el 2006.
h) Sí, ya que a cada uno de los valores que puede tomar t le corresponde uno y sólo un valor de W.
i) Sí, para indicar que t es función de W se puede escribir $t = H(W)$.
j) Sí.
k) Físicamente, suele considerarse que W está determinada por t y no que t está determinada por W. Por lo tanto, t es la variable dependiente y W es la variable independiente. Sin embargo, matemáticamente, en algunos casos, cualquiera de las dos variables puede considerarse como la variable independiente y la otra variable como la variable dependiente. La variable independiente es a la que se le pueden asignar diversos valores, y la otra variable cuyos valores dependen de los valores asignados es la variable dependiente.

1.16 Una variable Y está determinada por otra variable X de acuerdo con la ecuación $Y = 2X - 3$, donde el 2 y el 3 son exactos.

a) Encontrar Y para $X = 3, -2$ y 1.5.
b) Construir una tabla en la que se den los valores de Y para $X = -2, -1, 0, 1, 2, 3$ y 4.
c) Si $Y = F(X)$ denota que Y depende de X, determinar $F(2.4)$ y $F(0.8)$.
d) ¿Cuál es el valor de X que corresponde a $Y = 15$?
e) ¿Puede expresarse X como función de Y?

f) ¿Es Y una función univaluada de X?
g) ¿Es X una función univaluada de Y?

SOLUCIÓN

a) Para $X = 3$, $Y = 2X - 3 = 2(3) - 3 = 6 - 3 = 3$. Para $X = -2$, $Y = 2X - 3 = 2(-2) - 3 = -4 - 3 = -7$. Para $X = 1.5$, $Y = 2X - 3 = 2(1.5) - 3 = 3 - 3 = 0$.

b) En la tabla 1.2 se presentan los valores de Y obtenidos en el inciso *a*). Obsérvese que se pueden construir muchas tablas usando otros valores de X. La relación expresada por $Y = 2X - 3$ es equivalente a la colección de *todas* esas tablas.

Tabla 1.2

X	-2	-1	0	1	2	3	4
Y	-7	-5	-3	-1	1	3	5

c) $F(2.4) = 2(2.4) - 3 = 4.8 - 3 = 1.8$ y $F(0.8) = 2(0.8) - 3 = 1.6 - 3 = -1.4$.

d) En $Y = 2X - 3$ se sustituye $Y = 15$. Esto da $15 = 2X - 3$, $2X = 18$ y $X = 9$.

e) Sí. Ya que $Y = 2X - 3$, $Y + 3 = 2X$ y $X = \frac{1}{2}(Y + 3)$. Así, X queda expresada *explícitamente* como función de Y.

f) Sí. Ya que para cada uno de los valores que puede tomar X (que es una cantidad infinita) hay uno y sólo un valor de Y.

g) Sí. Ya que de acuerdo con el inciso *e*) $X = \frac{1}{2}(Y + 3)$, de manera que para cada uno de los valores que puede tomar Y hay uno y sólo un valor de X.

1.17 Si $Z = 16 + 4X - 3Y$, hallar el valor de Z que corresponda a: *a*) $X = 2$, $Y = 5$; *b*) $X = -3$, $Y = -7$; *c*) $X = -4$, $Y = 2$.

SOLUCIÓN

a) $Z = 16 + 4(2) - 3(5) = 16 + 8 - 15 = 9$

b) $Z = 16 + 4(-3) - 3(-7) = 16 - 12 + 21 = 25$

c) $Z = 16 + 4(-4) - 3(2) = 16 - 16 - 6 = -6$

A valores dados de X y Y, les corresponde un valor de Z. Para denotar que Z depende de X y de Y se escribe $Z = F(X, Y)$ (que se lee "Z es función de X y Y"). $F(2, 5)$ denota el valor de Z para $X = 2$ y $Y = 5$ que, de acuerdo con el inciso *a*), es 9. De igual manera, $F(-3, -7) = 25$ y $F(-4, 2) = -6$, de acuerdo con los incisos *b*) y *c*), respectivamente.

Las variables X y Y son las *variables independientes* y la variable Z es la *variable dependiente*.

1.18 Los gastos fijos de una empresa son de \$1 000 por día y los costos de producción de cada artículo son de \$25.

a) Escribir una ecuación que exprese el costo total de producción de x unidades por día.
b) Usando EXCEL, elaborar una tabla en la que se den los costos de producción de 5, 10, 15, 20, 25, 30, 35, 40, 45 y 50 unidades por día.
c) Evaluar e interpretar $f(100)$.

SOLUCIÓN

a) $f(x) = 1\,000 + 25x$.

b) Los números 5, 10, . . . , 50 se ingresan en B1:K1, la expresión = 1 000 + 25*B1 se ingresa en B2, se da clic y se arrastra desde B2 hasta K2 para obtener el resultado siguiente:

x	5	10	15	20	25	30	35	40	45
f(x)	1 025	1 050	1 075	1 100	1 125	1 150	1 175	1 200	1 225

c) $f(100) = 1\,000 + 25(100) = 1\,000 + 2\,500 = 3\,500$. Fabricar $x = 100$ unidades en un día cuesta 3 500.

1.19 El ancho de un rectángulo es x y el largo es $x + 10$.

a) Escribir una función, $A(x)$, que exprese el área en función de x.
b) Usar EXCEL para elaborar una tabla que dé el valor de $A(x)$ para $x = 0, 1, \ldots, 5$.
c) Escribir una función, $P(x)$, que exprese el perímetro en función de x.
d) Usar EXCEL para elaborar una tabla que dé el valor de $P(x)$ para $x = 0, 1, \ldots, 5$.

SOLUCIÓN

a) $A(x) = x(x + 10) = x^2 + 10x$

b) En las celdas B1:G1 se ingresan los números 0, 1, 2, 3, 4 y 5; en la celda B2 se ingresa la expresión =B1^2+10*B1, se da clic y se arrastra desde B2 hasta G2 con lo que se obtiene:

X	0	1	2	3	4	5
A(x)	0	11	24	39	56	75

c) $P(x) = x + (x + 10) + x + (x + 10) = 4x + 20$.

d) En las celdas B1:G1 se ingresan los números 0, 1, 2, 3, 4 y 5; en la celda B2 se ingresa la expresión =4*B1+20, se da clic y se arrastra desde B2 hasta G2 con lo que se obtiene:

X	0	1	2	3	4	5
P(x)	20	24	28	32	36	40

1.20 En un sistema de coordenadas rectangulares localizar los puntos que tienen como coordenadas: *a*) (5, 2), *b*) (2, 5), *c*) (−5, 1), *d*) (1, −3), *e*) (3, −4), *f*) (−2.5, −4.8), *g*) (0, −2.5) y *h*) (4, 0). Usar MAPLE para graficar estos puntos.

SOLUCIÓN

Véase la figura 1-2. A continuación se da el comando de MAPLE para graficar estos ocho puntos. Cada punto está representado por un círculo.

$L :=$ [[5, 2], [2, 5], [−5, 1], [1, −3], [3, −4], [−2.5, −4.8], [0, −2.5], [4, 0]];

pointplot (*L*, *font* = [*TIMES*, *BOLD*, *14*], *symbol* = *circle*);

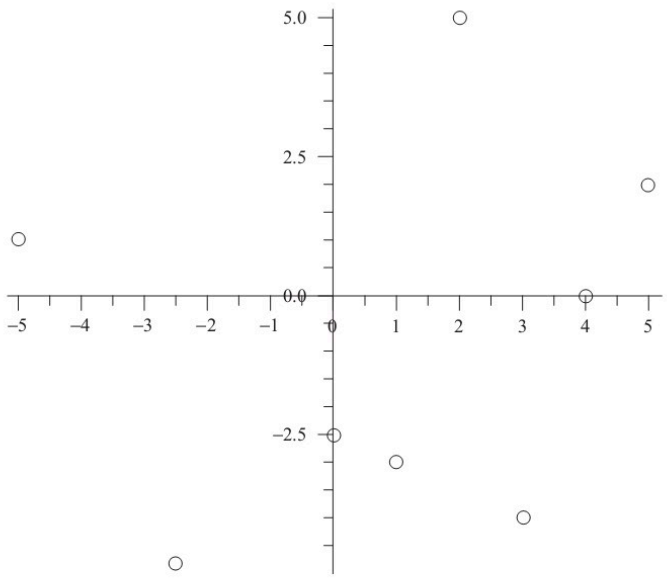


Figura 1-2 Gráfica MAPLE de puntos.

1.21 Graficar la ecuación $Y = 4X - 4$ usando MINITAB.

SOLUCIÓN

Obsérvese que la gráfica se extiende indefinidamente tanto en dirección positiva como en dirección negativa del eje X. Aquí se decidió, arbitrariamente, graficar sólo desde -5 hasta 5. En la figura 1-3 se muestra el diagrama de la recta $Y = 4X - 4$ obtenida con MINITAB. De la barra de herramientas se selecciona la secuencia "**Graph ⇒ Scatterplots**" para activar scatter plots (gráfica de dispersión). Los puntos sobre la recta se obtienen ingresando los enteros desde -5 hasta 5 y usando la calculadora de MINITAB para calcular los valores correspondientes de Y. Los valores de X y Y son los siguientes:

X	−5	−4	−3	−2	−1	0	1	2	3	4	5
Y	−24	−20	−16	−12	−8	−4	0	4	8	12	16

Los puntos se han unido para dar una idea de cómo se ve la gráfica de la ecuación $Y = 4X - 4$.

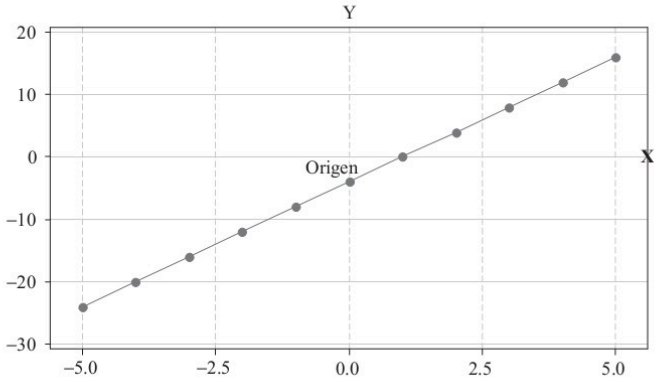


Figura 1-3 Gráfica MINITAB de una función lineal.

1.22 Grafique la ecuación $Y = 2X^2 - 3X - 9$ usando EXCEL.

SOLUCIÓN

Tabla 1.3 Valores de una función cuadrática generados con EXCEL

X	−5	−4	−3	−2	−1	0	1	2	3	4	5
Y	56	35	18	5	−4	−9	−10	−7	0	11	26

Se usó EXCEL para elaborar esta tabla que da los valores de Y para los valores de X igual a –5, $-4, \ldots, 5$. Se ingresa la expresión =2*B1^2-3*B1-9 en la celda B2, se da clic y se arrastra desde B2 hasta L2. Para obtener la gráfica que se muestra en la figura 1-4 se usa el asistente para gráficos de EXCEL. Ésta es una *función cuadrática*. Las *raíces* (puntos en los que la gráfica cruza el eje x) de esta función cuadrática están una en $X = 3$ y la otra entre –2 y –1. Haciendo clic sobre el asistente para gráficos de EXCEL, se muestran las diversas gráficas que es posible hacer. Obsérvese que a medida que X toma valores cada vez más grandes, tanto positivos como negativos, la gráfica de esta función cuadrática va hacia el infinito positivo. Obsérvese también que la gráfica toma su valor más bajo cuando X está entre 0 y 1.

1.23 La tabla 1.4 muestra el aumento de la cantidad de diabéticos desde 1997 hasta 2005. Grafique estos datos.

Tabla 1.4 Cantidad de nuevos diabéticos

Año	1977	1998	1999	2000	2001	2002	2003	2004	2005
Millones	0.88	0.90	1.01	1.10	1.20	1.25	1.28	1.36	1.41

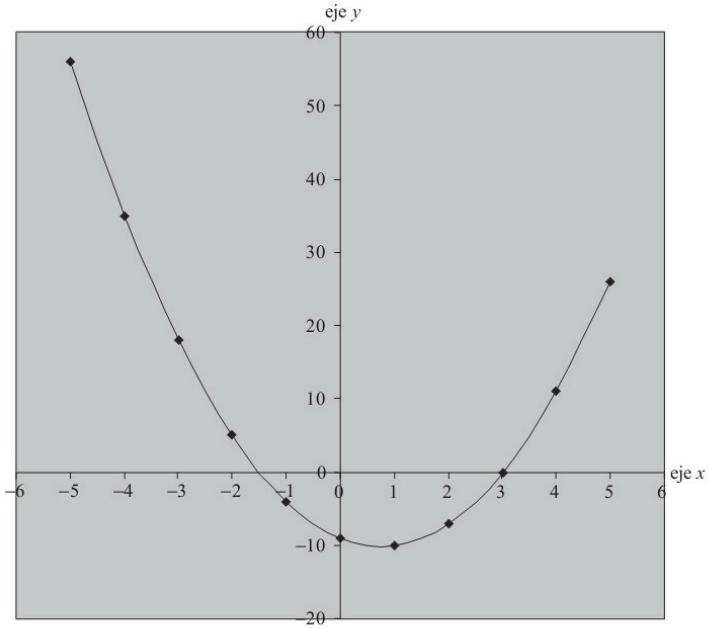


Figura 1-4 Diagrama EXCEL de una curva llamada parábola.

SOLUCIÓN

Primer método

La primer gráfica que se muestra en la figura 1-5 es la *gráfica de una serie de tiempos*. En este diagrama se presentan los nuevos casos de diabetes desde 1997 hasta 2005. Se muestra que durante este periodo la cantidad de nuevos casos ha ido aumentando.

Segundo método

A la figura 1-6 se le conoce como *gráfica de barras*, *carta de barras* o *diagrama de barras*. El ancho de las barras, que en todas es el mismo, no tiene ningún significado en este caso y pueden ser de cualquier tamaño en tanto no se traslapen.

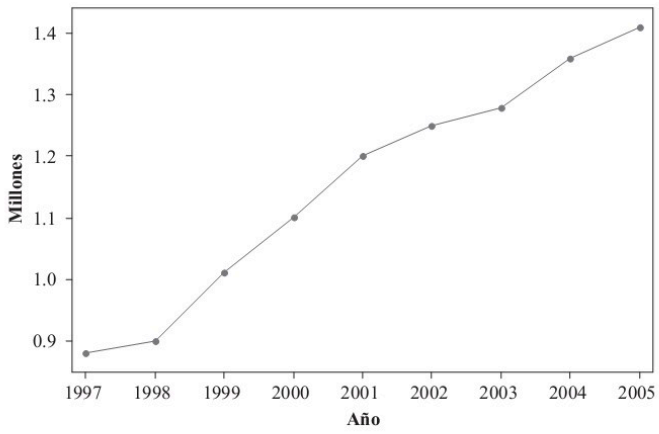


Figura 1-5 MINITAB, serie de tiempos de nuevos casos de diabetes por año.

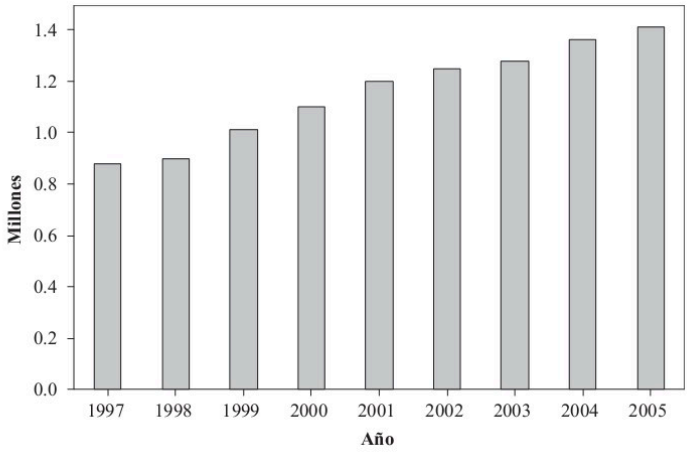


Figura 1-6 MINITAB, gráfica de barras de los nuevos casos de diabetes por año.

Tercer método

En la figura 1-7 se muestra una gráfica de barras en la que las barras son horizontales en vez de verticales.

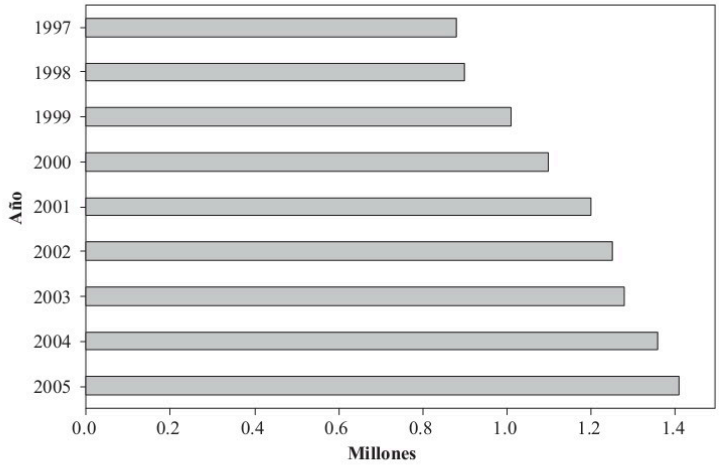


Figura 1-7 MINITAB, gráfica de barras horizontales de nuevos casos de diabetes por año.

1.24 Grafique los datos del problema 1.14 usando una gráfica de MINITAB para serie de tiempos, una gráfica de barras agrupadas con efecto tridimensional (3-D) de EXCEL y una gráfica de barras apiladas con efecto 3-D de EXCEL.

SOLUCIÓN

Las soluciones se dan en las figuras 1-8, 1-9 y 1-10.

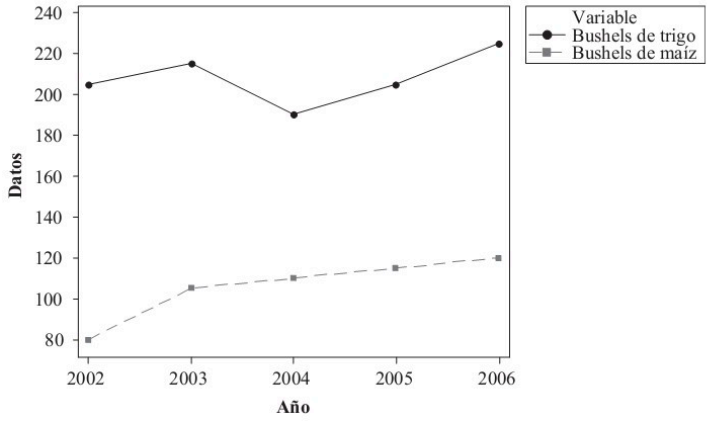


Figura 1-8 MINITAB, serie de tiempos de la producción (2002 a 2006) de trigo y maíz.

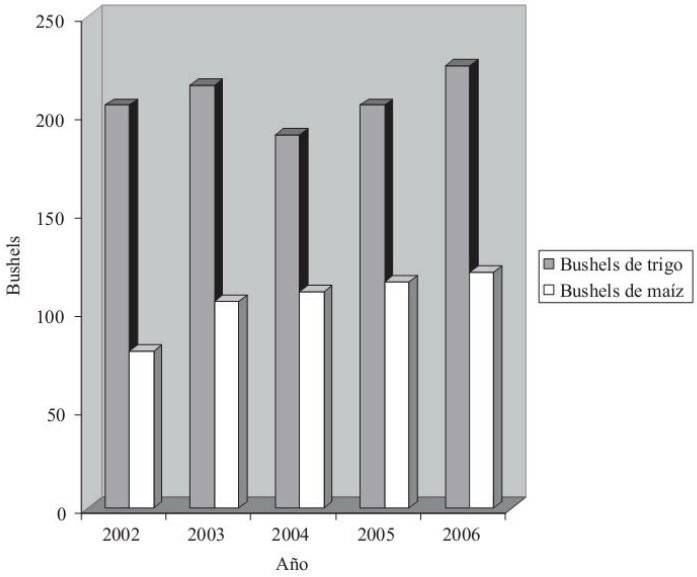


Figura 1-9 EXCEL, barras agrupadas con efecto 3-D.

1.25 *a*) Expresar las cantidades anuales de bushels de trigo y de maíz, presentadas en la tabla 1.1 del problema 1.4, como porcentajes de la producción anual total.

b) Graficar los porcentajes obtenidos en el inciso *a*).

SOLUCIÓN

a) El porcentaje de trigo correspondiente al 2002 es $= 205/(205 + 80) = 71.9\%$ y porcentaje de maíz $= 100\% - 71.9\%$ $= 28.1\%$, etc. Estos porcentajes se muestran en la tabla 1.5.

b) Las *columnas apiladas 100%* comparan los porcentajes con la contribución de cada valor al total de cada categoría (figura 1-11).

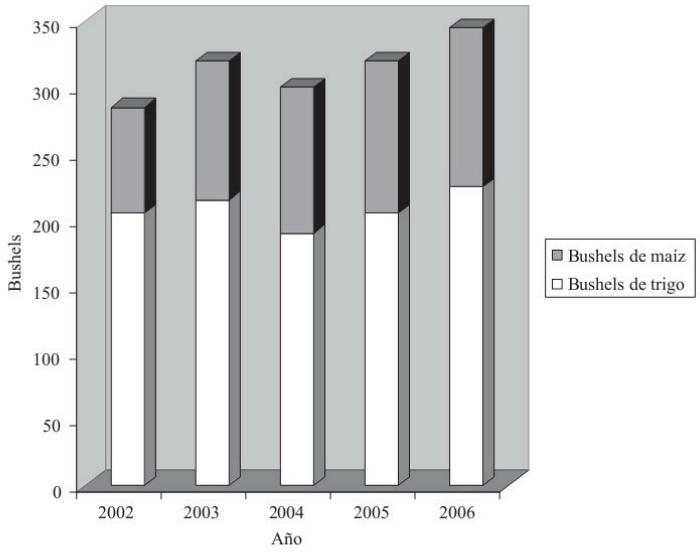


Figura 1-10 EXCEL, barras apiladas con efecto 3-D.

Tabla 1.5 Producción de trigo y maíz desde 2002 hasta 2006

Año	Trigo (%)	Maíz (%)
2002	71.9	28.1
2003	67.2	32.8
2004	63.3	36.7
2005	64.1	35.9
2006	65.2	34.8

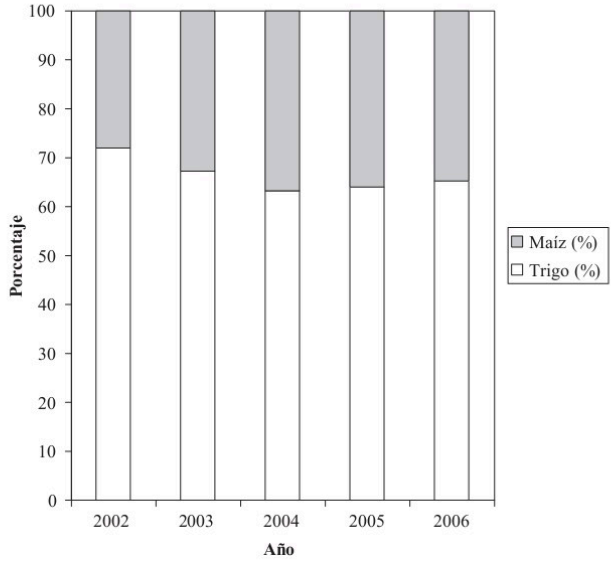


Figura 1-11 EXCEL, columnas 100% apiladas.

1.26 En un número reciente de *USA Today*, una nota titulada "Peligro en línea", informa de un estudio realizado en 1 500 niños entre 10 y 17 años de edad. Presentar la información de la tabla 1.6 en una gráfica de barras agrupadas y en una gráfica de barras apiladas.

Tabla 1.6

	Prostitución	Contacto con la pornografía	Acoso
2000	19%	25%	6%
2005	13%	34%	9%

SOLUCIÓN

En la figura 1-12 se muestra la gráfica de barras con columnas agrupadas y en la figura 1-13 la gráfica de barras con columnas apiladas obtenida con esta información.

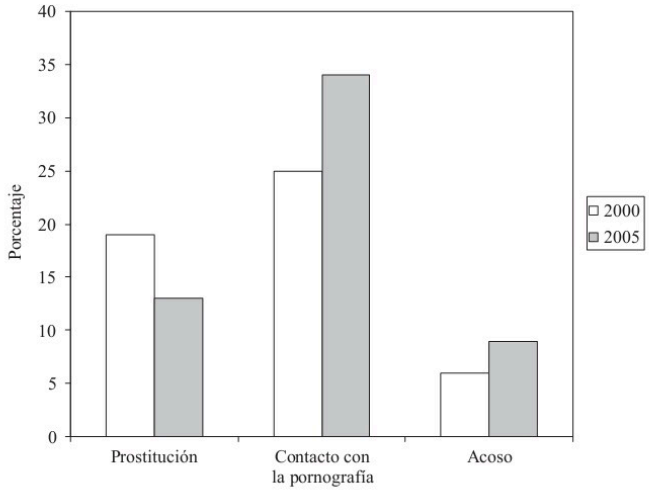


Figura 1-12 EXCEL, gráfica de barras con columnas agrupadas.

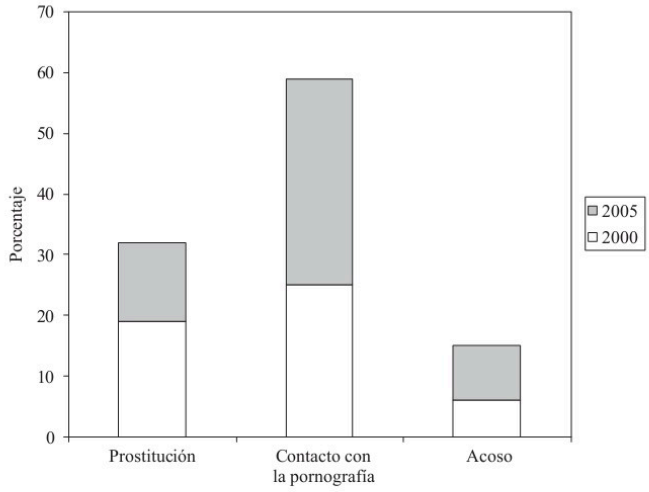


Figura 1-13 EXCEL, gráfica de barras con columnas apiladas.

1.27 En una nota reciente de *USA Today* titulada "¿Dónde están los estudiantes universitarios?", se informó que en Estados Unidos hay más de 17.5 millones de universitarios que estudian en más de 6 400 escuelas. En la tabla 1.7 se da la matrícula de acuerdo al tipo de escuela.

Tabla 1.7 ¿Dónde están los estudiantes universitarios?

Tipo de escuela	Porcentaje
Pública de 2 años	43
Pública de 4 años	32
Privada no lucrativa de 4 años	15
Privada de 2 y 4 años	6
Privada de menos de 4 años	3
Otras	1

Con la información de la tabla 1.7 construya una gráfica de barras 3-D usando EXCEL y una gráfica de barras usando MINITAB.

SOLUCIÓN

Las figuras 1-14 y 1-15 dan las gráficas pedidas.

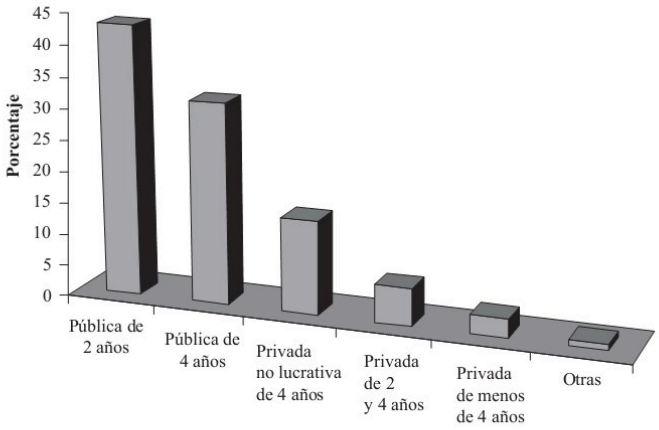


Figura 1-14 EXCEL, gráfica de barras 3-D con los datos de la tabla 1.7.

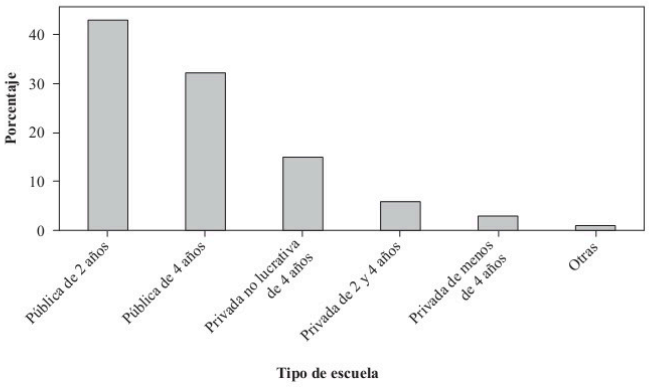


Figura 1-15 MINITAB, gráfica de barras con los datos de la tabla 1.7.

1.28 Los estadounidenses tienen en promedio 2.8 televisores por hogar. Con los datos de la tabla 1.8 elabore una gráfica de pastel usando EXCEL.

Tabla 1.8 Televisores por hogar

Televisores	Porcentaje
Ninguno	2
Uno	15
Dos	29
Tres	26
Cuatro	16
Más de cinco	12

SOLUCIÓN

En la figura 1-16 se presenta la gráfica de pastel obtenida con EXCEL para los datos de la tabla 1.8.

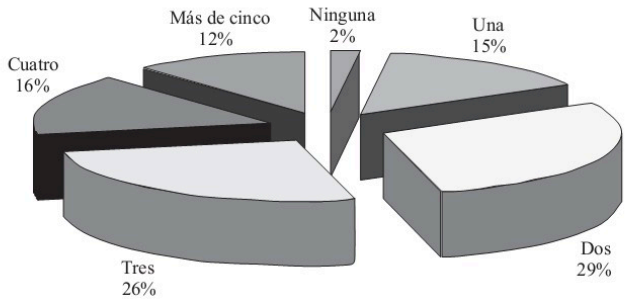


Figura 1-16 EXCEL, gráfica de pastel con la información de la tabla 1.8.

ECUACIONES

1.29 Resuelva las siguientes ecuaciones:

a) $4a - 20 = 8$ *c*) $18 - 5b = 3(b + 8) + 10$

b) $3X + 4 = 24 - 2X$ *d*) $\dfrac{Y + 2}{3} + 1 = \dfrac{Y}{2}$

SOLUCIÓN

a) Sumar 20 a ambos miembros: $4a - 20 + 20 = 8 + 20$ o bien $4a = 28$.
Dividir ambos lados entre 4: $4a/4 = 28/4$ y $a = 7$.
Verificación: $4(7) - 20 = 8$, $28 - 20 = 8$ y $8 = 8$.

b) Restar 4 de ambos miembros: $3X + 4 - 4 = 24 - 2X - 4$ o bien $3X = 20 - 2X$.
Sumar $2X$ a ambos lados: $3X + 2X = 20 - 2X + 2X$ o bien $5X = 20$.
Dividir ambos lados entre 5: $5X/5 = 20/5$ y $X = 4$.
Verificación: $3(4) + 4 = 24 - 2(4)$, $12 + 4 = 24 - 8$ y $16 = 16$.

Este resultado se puede obtener mucho más rápidamente si se observa que todos los términos se pueden pasar o trasponer de un miembro a otro de la ecuación cambiándoles simplemente el signo. Así, se puede escribir

$$3X + 4 = 24 - 2X \qquad 3X + 2X = 24 - 4 \qquad 5X = 20 \qquad X = 4$$

c) $18 - 5b = 3b + 24 + 10$ y $18 - 5b = 34$.
Transponiendo, $-5b - 3b = 34 - 18$ o bien $-8b = 16$.
Dividiendo entre -8, $-8b/(-8) = 16/(-8)$ y $b = -2$.
Verificación: $18 - 5(-2) = 3(-2 + 8) + 10$, $18 + 10 = 3(6) + 10$ y $28 = 28$.

d) Primero se multiplican ambos miembros por 6, que es el mínimo común denominador.

$$6\left(\frac{Y+2}{3}+1\right) = 6\left(\frac{Y}{2}\right) \qquad 6\left(\frac{Y+2}{3}\right) + 6(1) = \frac{6Y}{2} \qquad 2(Y+2) + 6 = 3Y$$

$$2Y + 4 + 6 = 3Y \qquad 2Y + 10 = 3Y \qquad 10 = 3Y - 2Y \qquad Y = 10$$

Verificación: $\frac{10+2}{3} + 1 = \frac{10}{2}$, $\frac{12}{3} + 1 = \frac{10}{2}$, $4 + 1 = 5$ y $5 = 5$.

1.30 Resolver los siguientes sistemas de ecuaciones simultáneas:

a) $3a - 2b = 11$, $5a + 7b = 39$ *b*) $5X + 14Y = 78$, $7X + 3Y = -7$ *c*) $3a + 2b + 5c = 15$, $7a - 3b + 2c = 52$, $5a + b - 4c = 2$

SOLUCIÓN

a) Multiplicando la primera ecuación por 7: $21a - 14b = 77$ (*1*)
Multiplicando la segunda ecuación por 2: $10a + 14b = 78$ (*2*)
Sumando: $31a = 155$
Dividiendo entre 31: $a = 5$

Obsérvese que multiplicando cada una de las ecuaciones dadas por un número adecuado, se obtienen las ecuaciones equivalentes (*1*) y (*2*), en las que los coeficientes de la variable *b* son numéricamente iguales. Después, sumando las dos ecuaciones se elimina la incógnita *b* y se encuentra *a*.

Sustituyendo $a = 5$ en la primera ecuación: $3(5) - 2b = 11$, $-2b = -4$ y $b = 2$. Por lo tanto, $a = 5$ y $b = 2$.

Verificación: $3(5) - 2(2) = 11$, $15 - 4 = 11$ y $11 = 11$; $5(5) + 7(2) = 39$, $25 + 14 = 39$ y $39 = 39$.

b) Multiplicando la primera ecuación por 3: $15X + 42Y = 234$ (*3*)
Multiplicando la segunda ecuación por -14: $-98X - 42Y = 98$ (*4*)
Sumando: $-83X = 332$
Dividiendo entre -83: $X = -4$

Sustituyendo $X = -4$ en la primera ecuación: $5(-4) + 14Y = 78$, $14Y = 98$, y $Y = 7$.
Por lo tanto, $X = -4$ y $Y = 7$.
Verificación: $5(-4) + 14(7) = 78$, $-20 + 98 = 78$ y $78 = 78$; $7(-4) + 3(7) = -7$, $-28 + 21 = -7$ y $-7 = -7$.

c) Multiplicando la primera ecuación por 2: $6a + 4b + 10c = 30$
Multiplicando la segunda ecuación por -5: $-35a + 15b - 10c = -260$
Sumando: $-29a + 19b = -230$ (*5*)

Multiplicando la segunda ecuación por 2: $14a - 6b + 4c = 104$
Repitiendo la tercera ecuación: $5a + b - 4c = 2$
Sumando: $19a - 5b = 106$ (*6*)

De esta manera se ha eliminado *c* y quedan dos ecuaciones (*5*) y (*6*), que deben resolverse simultáneamente para encontrar *a* y *b*.

Multiplicando la ecuación (*5*) por 5: $-145a + 95b = -1150$
Multiplicando la ecuación (*6*) por 19: $361a - 95b = 2014$
Sumando: $216a = 864$
Dividiendo entre 216: $a = 4$

Sustituyendo $a = 4$ en la ecuación (*5*) o bien (*6*), se encuentra que $b = -6$.

Sustituyendo $a = 4$ y $b = -6$ en cualquiera de las ecuaciones dadas, se obtiene $c = 3$.

Por lo tanto, $a = 4$, $b = -6$ y $c = 3$.
Verificación: $3(4) + 2(-6) + 5(3) = 15$ y $15 = 15$; $7(4) - 3(-6) + 2(3) = 52$ y $52 = 52$; $5(4) + (-6) - 4(3) = 2$ y $2 = 2$.

DESIGUALDADES

1.31 Expresar con palabras el significado de:

a) $N > 30$ *b*) $X \leq 12$ *c*) $0 < p \leq 1$ *d*) $\mu - 2t < X < \mu + 2t$

SOLUCIÓN

a) N es mayor que 30.
b) X es menor o igual a 12.
c) p es mayor que cero y menor o igual a 1.
d) X es mayor que $\mu - 2t$ pero menor que $\mu + 2t$.

1.32 Traducir a símbolos lo siguiente:

a) La variable X toma valores entre 2 y 5 inclusive.
b) La media aritmética $\bar{X}$ es mayor que 28.42 y menor que 31.56.
c) m es un número positivo menor o igual a 10.
d) P es un número no negativo.

SOLUCIÓN

a) $2 \leq X \leq 5$; *b*) $28.42 < \bar{X} < 31.56$; *c*) $0 < m \leq 10$; *d*) $P \geq 0$.

1.33 Empleando los signos de desigualdad, ordenar los números 3.42, –0.6, –2.1, 1.45 y –3 en *a*) en orden creciente de magnitud y en *b*) en orden decreciente de magnitud.

SOLUCIÓN

a) $-3 < -2.1 < -0.6 < 1.45 < 3.42$
b) $3.42 > 1.45 > -0.6 > -2.1 > -3$

Obsérvese que cuando estos puntos se grafican como puntos en la línea (ver problema 1.18), aumentan de izquierda a derecha.

1.34 Resolver cada una de las desigualdades siguientes (es decir, despejar X):

a) $2X < 6$ *c*) $6 - 4X < -2$ *e*) $-1 \leq \dfrac{3 - 2X}{5} \leq 7$

b) $3X - 8 \geq 4$ *d*) $-3 < \dfrac{X - 5}{2} < 3$

SOLUCIÓN

a) Dividiendo ambos lados entre 2 se obtiene $X < 3$.
b) Sumando 8 a ambos lados, $3X \geq 12$; dividiendo ambos lados entre 3, $X \geq 4$.
c) Sumando -6 a ambos lados, $-4X < -8$; dividiendo ambos lados entre -4, $X > 2$. Obsérvese que como ocurre en las ecuaciones, también en una desigualdad se puede transponer un término de un lado a otro de la desigualdad cambiando simplemente el signo del término; por ejemplo, en el inciso *b*), $3X \geq 8 + 4$.
d) Multiplicando por 2, $-6 < X - 5 < 6$; sumando 5, $-1 < X < 11$.
e) Multiplicando por 5, $-5 \leq 3 - 2X \leq 35$; sumando -3, $-8 \leq -2X \leq 32$; dividiendo entre -2, $4 \geq X \geq -16$, o bien $-16 \leq X \leq 4$.

LOGARITMOS Y PROPIEDADES DE LOS LOGARITMOS

1.35 Utilizar la definición $y = \log_b x$ para hallar los logaritmos siguientes y después usar EXCEL para verificar la respuesta. (Obsérvese que $y = \log_b x$ significa que $b^y = x$.

a) Encontrar el log de base 2 de 32.
b) Encontrar el log de base 4 de 64.
c) Encontrar el log de base 6 de 216.
d) Encontrar el log de base 8 de 4 096.
e) Encontrar el log de base 10 de 10 000.

SOLUCIÓN

a) 5; *b*) 3; *c*) 3; *d*) 4; *e*) 4.

La expresión de EXCEL =LOG(32,2) da 5, =LOG(64,4) da 3, =LOG(216,6) da 3, =LOG(4 096,8) da 4 y =LOG(10 000, 10) da 4.

1.36 Empleando las propiedades de los logaritmos, volver a escribir los logaritmos siguientes como sumas y diferencias de logaritmos.

a) $\ln\left(\dfrac{x^2y^3z}{ab}\right)$ *b*) $\log\left(\dfrac{a^2b^3c}{yz}\right)$

Empleando las propiedades de los logaritmos, reescribir los logaritmos siguientes como un solo logaritmo.

c) $\ln(5) + \ln(10) - 2\ln(5)$ *d*) $2\log(5) - 3\log(5) + 5\log(5)$

SOLUCIÓN

a) $2\ln(x) + 3\ln(y) + \ln(z) - \ln(a) - \ln(b)$
b) $2\log(a) + 3\log(b) + \log(c) - \log(y) - \log(z)$
c) $\ln(2)$
d) $\log(625)$

1.37 Usando SAS y SPSS, graficar $y = \ln(x)$.

SOLUCIÓN

Las soluciones se muestran en las figuras 1-17 y 1-18.

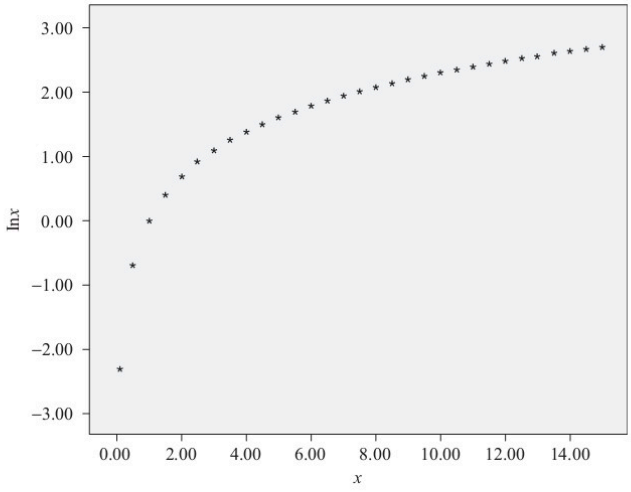


Figura 1-17 Gráfica SPSS de $y = \ln(x)$.

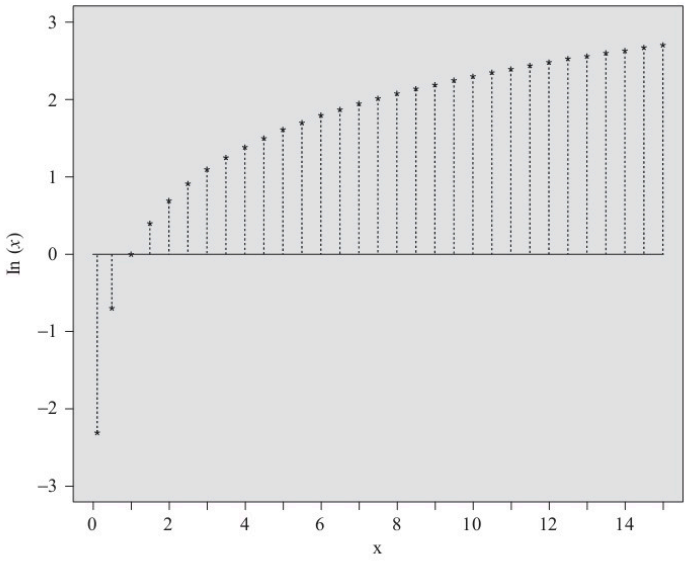


Figura 1-18 Gráfica SAS de $y = \ln(x)$.

Las figuras 1-17 y 1-18 muestran una gráfica de la curva $y = \ln(x)$. A medida que x se aproxima a 0, los valores de $\ln(x)$ se aproximan cada vez más a $-\infty$. A medida que x crece, los valores de $\ln(x)$ se aproximan a $+\infty$.

ECUACIONES LOGARÍTMICAS

1.38 Resolver la ecuación logarítmica $\ln(x) = 10$.

SOLUCIÓN

Empleando la definición de logaritmo, $x = e^{10} = 22026.47$. Como verificación se saca el logaritmo natural de 22026.47 y se obtiene 10.00000019.

1.39 Resolver la ecuación logarítmica $\log(x + 2) + \log(x - 2) = \log(5)$.

SOLUCIÓN

El lado izquierdo se puede escribir como $\log[(x + 2)(x - 2)]$. Se obtiene la ecuación $\log(x + 2)(x - 2) = \log(5)$, de la cual $(x + 2)(x - 2) = (5)$. A partir de la cual sigue la ecuación $x^2 - 4 = 5$ o bien $x^2 = 9$ o bien $x = -3$ o bien 3. Cuando estos valores se verifican en la ecuación original, $x = -3$ debe descartarse como solución porque el logaritmo de números negativos no está definido. Si en la ecuación original se sustituye $x = 3$, se tiene $\log(5) + \log(1) = \log(5)$, ya que $\log(1) = 0$.

1.40 Resuelva la ecuación logarítmica $\log(a + 4) - \log(a - 2) = 1$.

SOLUCIÓN

Esta ecuación se puede escribir como $\log((a + 4)/(a - 2)) = 1$. Aplicando la definición de logaritmo, se tiene $(a + 4)/(a - 2) = 10^1$ o bien $a + 4 = 10a - 20$. Despejando a, $a = 24/9 = 2.6$ (siendo el 6 periódico). Sustituyendo en la ecuación original a por 2.6667, se tiene $0.8239 - (-0.1761) = 1$. La única solución es 2.6667.

1.41 Resolver la ecuación logarítmica $\ln(x)^2 - 1 = 0$.

SOLUCIÓN

Esta ecuación se puede factorizar como $[\ln(x) + 1][\ln(x) - 1] = 0$. Haciendo el factor $\ln(x) + 1 = 0$, se obtiene $\ln(x) = -1$ o bien $x = e^{-1} = 0.3678$. Haciendo el segundo factor $\ln(x) - 1 = 0$, se tiene $\ln(x) = 1$ o bien $x = e^1 = 2.7183$. Ambos valores son solución de la ecuación.

1.42 En la ecuación logarítmica siguiente, despejar x: $2\log(x + 1) - 3\log(x + 1) = 2$.

SOLUCIÓN

Esta ecuación se puede escribir como $\log[(x + 1)^2/(x + 1)^3] = 2$ o bien $\log[1/(x + 1)] = 2$ o bien $\log(1) - \log(x + 1)] = 2$ o bien $0 - \log(x + 1) = 2$ o bien $\log(x + 1) = -2$ o bien $x + 1 = 10^{-2}$ o bien $x = -0.99$. Sustituyendo en la ecuación original, se encuentra $2\log(0.01) - 3\log(0.01) = 2$. Por lo tanto, la solución satisface la ecuación.

1.43 Para resolver ecuaciones logarítmicas que no son fáciles de resolver a mano, se puede usar el paquete de software MAPLE. Resolver la ecuación siguiente usando MAPLE.

$$\log(x + 2) - \ln(x^2) = 4$$

SOLUCIÓN

El comando de MAPLE para resolver la ecuación es "solve(log10(x + 2) − ln(x^2) = 4);" la solución dada es −0.154594. Obsérvese que MAPLE usa log10 para el logaritmo común.

Para comprobar que la solución es correcta, sustituyendo en la ecuación original se tiene log(1.845406) − ln(0.023899) que es igual a 4.00001059.

1.44 EXCEL también se puede usar para resolver ecuaciones logarítmicas. Resolver la siguiente ecuación logarítmica usando EXCEL: $\log(x + 4) + \ln(x + 5) = 1$.

SOLUCIÓN

En la figura 1-19 se da la hoja de cálculo de EXCEL.

–3	–0.30685	LOG10(A1+4)+LN(A1+5)^−1
–2	0.399642	
–1	0.863416	
0	1.211498	
1	1.490729	
2	1.724061	
3	1.92454	
4	2.100315	
5	2.256828	

–3	–0.30685	LOG10(A11+4)+LN(A11+5)^−1
–2.9	–0.21667	
–2.8	–0.13236	
–2.7	–0.05315	
–2.6	0.021597	
–2.5	0.092382	
–2.4	0.159631	
–2.3	0.223701	
–2.2	0.284892	
–2.1	0.343464	
–2	0.399642	

Figura 1-19 EXCEL, hoja de trabajo para el problema 1.44

Se puede usar la técnica iterativa que se muestra antes. La mitad superior encuentra que la raíz de $\log(x+4) + \ln(x+5) - 1$ está entre -3 y -2. La mitad inferior encuentra que la raíz está entre -2.7 y -2.6. Para dar la raíz con la exactitud que se desee, sólo hace falta continuar con este proceso. Al usar esta técnica se emplea la de clic y arrastre.

1.45 Encuentre la solución al problema 1.44 usando MAPLE.

SOLUCIÓN

El comando de MAPLE "> **solve(log 10(x + 4) + ln(x + 5) = 1)**;" da como solución -2.62947285. Compare este resultado con el obtenido en el problema 1.44.

PROBLEMAS SUPLEMENTARIOS

VARIABLES

1.46 Cuáles de los datos siguientes son datos discretos y cuáles son datos continuos.

a) Precipitación pluvial, en pulgadas, en una ciudad, en diversos meses del año.
b) Velocidad de un automóvil, en millas por hora.
c) Cantidad de billetes de $20 que circulan en Estados Unidos en determinado momento.
d) Valor total diario de las acciones vendidas en la bolsa.
e) Cantidad de estudiantes inscritos anualmente en una universidad.

1.47 Dar el dominio de cada una de las variables siguientes e indicar si es una variable discreta o continua.

a) Cantidad anual W de bushels de trigo por acre que se producen en una granja.
b) Cantidad N de individuos en una familia.
c) Estado civil de un individuo.
d) Tiempo T de vuelo de un misil.
e) Número P de pétalos que tiene una flor.

REDONDEO DE CANTIDADES NUMÉRICAS, NOTACIÓN CIENTÍFICA Y CIFRAS SIGNIFICATIVAS

1.48 Redondear cada uno de los números siguientes como se indica.

a)	3 256	a la centena más cercana
b)	5.781	a la décima más cercana
c)	0.0045	a la milésima más cercana
d)	46.7385	a la centésima más cercana
e)	125.9995	a dos lugares decimales
f)	3 502 378	al millón más cercano
g)	148.475	a la unidad más cercana
h)	0.000098501	a la millonésima más cercana
i)	2 184.73	a la decena más cercana
j)	43.87500	a la centésima más cercana

1.49 Expresar cada número sin usar potencias de 10.

a)	132.5×10^4	*d*)	$7\,300 \times 10^6$
b)	418.72×10^{-5}	*e*)	3.487×10^{-4}
c)	280×10^{-7}	*f*)	0.0001850×10^5

1.50 ¿Cuántas cifras significativas hay en cada una de las cantidades siguientes entendiendo que se han registrado exactamente?

a) 2.54 cm
b) 0.004500 yd
c) 3 510 000 bu
d) 3.51 millones de bu
e) 10.000100 ft
f) 378 personas
g) 378 oz
h) 4.50×10^{-3} km
i) 500.8×10^{5} kg
j) 100.00 mi

1.51 ¿Cuál es el error máximo en cada una de las mediciones siguientes, entendiéndose que han sido registradas exactamente? En cada caso dar el número de cifras significativas.

a) 7.20 millones bu
b) 0.00004835 cm
c) 5 280 ft
d) 3.5×10^{8} m
e) 186 000 mi/s
f) 186 mil mi/s

1.52 Escribir cada uno de los números siguientes en notación científica. Supóngase que todas las cifras son significativas a menos que se indique otra cosa.

a) 0.000317
b) 428 000 000 (cuatro cifras significativas)
c) 21 600.00
d) 0.000009810
e) 732 mil
f) 18.0 diezmilésimas

CÁLCULOS

1.53 Mostrar que: *a*) el producto y *b*) el cociente de los números 72.48 y 5.16, considerando que tienen cuatro y tres cifras significativas, respectivamente, no puede ser exacto a más de tres cifras significativas. Escribir el producto y el cociente exactos.

1.54 Realizar cada una de las operaciones indicadas. A menos que se indique otra cosa, supóngase que los números se han registrado exactamente.

a) 0.36×781.4

b) $\dfrac{873.00}{4.881}$

c) $5.78 \times 2\,700 \times 16.00$

d) $\dfrac{0.00480 \times 2\,300}{0.2084}$

e) $\sqrt{120 \times 0.5386 \times 0.4614}$ (120 exactos)

f) $\dfrac{(416\ 000)(0.000187)}{\sqrt{73.84}}$

g) $14.8641 + 4.48 - 8.168 + 0.36125$

h) $4\,173\,000 - 170\,264 + 1\,820\,470 - 78\,320$ (estos números son exactos a cuatro, seis, seis y cinco cifras significativas, respectivamente)

i) $\sqrt{\dfrac{7(4.386)^2 - 3(6.47)^2}{6}}$ (el 3, el 6 y el 7 son exactos)

j) $4.120\sqrt{\dfrac{3.1416[(9.483)^2 - (5.075)^2]}{0.0001980}}$

1.55 Evaluar cada una de las expresiones siguientes, si $U = -2$, $V = \frac{1}{2}$, $W = 3$, $X = -4$, $Y = 9$ y $Z = \frac{1}{6}$, donde se entiende que todos los números son exactos.

a) $4U + 6V - 2W$

b) $\dfrac{XYZ}{UVW}$

c) $\dfrac{2X - 3Y}{UW + XV}$

d) $3(U - X)^2 + Y$

e) $\sqrt{U^2 - 2UV + W}$

f) $3X(4Y + 3Z) - 2Y(6X - 5Z) - 25$

g) $\sqrt{\dfrac{(W-2)^2}{V}+\dfrac{(Y-5)^2}{Z}}$

h) $\dfrac{X-3}{\sqrt{(Y-4)^2+(U+5)^2}}$

i) X^3+5X^2-6X-8

j) $\dfrac{U-V}{\sqrt{U^2+V^2}}[U^2V(W+X)]$

FUNCIONES, TABLAS Y GRÁFICAS

1.56 Una variable Y está determinada por una variable X de acuerdo con la ecuación $Y = 10 - 4X$.

a) Encontrar el valor de Y para $X = -3, -2, -1, 0, 1, 2, 3, 4$ y 5 y presentar los resultados en una tabla.

b) Encontrar el valor de Y para $X = -2.4, -1.6, -0.8, 1.8, 2.7, 3.5$ y 4.6.

c) Si $Y = F(X)$ denota que Y depende de X, hallar $F(2.8)$, $F(-5)$, $F(\sqrt{2})$ y $F(-\pi)$.

d) Dar el valor de X que corresponde a $Y = -2, 6, -10, 1.6, 16, 0$ y 10.

e) Expresar X explícitamente como función de Y.

1.57 Si $Z = X^2 - Y^2$, encontrar el valor de Z para: *a*) $X = -2$, $Y = 3$ y *b*) $X = 1$, $Y = 5$. *c*) Si se usa la notación funcional $Z = F(X, Y)$, encontrar $F(-3, -1)$.

1.58 Si $W = 3XZ - 4Y^2 + 2XY$, encontrar el valor de W para: *a*) $X = 1$, $Y = -2$, $Z = 4$ y *b*) $X = -5$, $Y = -2$, $Z = 0$. *c*) Si se usa la notación funcional $W = F(X, Y, Z)$, encontrar $F(3, 1, -2)$.

1.59 En un sistema de coordenadas rectangulares, localizar los puntos cuyas coordenadas son: *a*) (3, 2), *b*) (2, 3), *c*) (−4, 4), *d*) (4, −4), *e*) (−3, −2), *f*) (−2, −3), *g*) (−4.5, 3), *h*) (−1.2, −2.4), *i*) (0, −3) y *j*) (1.8, 0).

1.60 Grafique las ecuaciones: *a*) $Y = 10 - 4X$ (ver problema 1.56), *b*) $Y = 2X + 5$, *c*) $Y = \frac{1}{3}(X - 6)$, *d*) $2X + 3Y = 12$ y *e*) $3X - 2Y = 6$.

1.61 Graficar las ecuaciones: *a*) $Y = 2X^2 + X - 10$ y *b*) $Y = 6 - 3X - X^2$.

1.62 Graficar $Y = X^3 - 4X^2 + 12X - 6$.

1.63 En la tabla 1.9 se presenta la cantidad de gimnasios y la cantidad de sus miembros en millones para los años desde 2000 hasta 2005. Emplear un paquete de software para trazar una gráfica de serie de tiempos para los gimnasios y otra para sus miembros.

Tabla 1.9

Año	2000	2001	2002	2003	2004	2005
Gimnasios	13 000	13 225	15 000	20 000	25 500	28 500
Miembros	32.5	35.0	36.5	39.0	41.0	41.3

1.64 Emplear un paquete de software para trazar, con los datos de la tabla 1.9, una gráfica de barras de los gimnasios y de los miembros.

1.65 Emplear EXCEL para trazar, con los datos de la tabla 1.9, un diagrama de dispersión de los gimnasios y de los miembros.

1.66 En la tabla 1.10 se da la mortalidad infantil por 1 000 nacidos vivos, para blancos y para no blancos, desde el año 2000 hasta el 2005. Usar una gráfica adecuada para representar estos datos.

Tabla 1.10

Año	2000	2001	2002	2003	2004	2005
Blancos	6.6	6.3	6.1	6.0	5.9	5.7
No blancos	7.6	7.5	7.3	7.2	7.1	6.8

1.67 En la tabla 1.11 se dan las velocidades orbitales de los planetas de nuestro sistema solar. Graficar estos datos.

Tabla 1.11

Planeta	Mercurio	Venus	Tierra	Marte	Júpiter	Saturno	Urano	Neptuno	Plutón
Velocidad (mi/s)	29.7	21.8	18.5	15.0	8.1	6.0	4.2	3.4	3.0

1.68 En la tabla 1.12 se da la matrícula (en miles) de las escuelas públicas en los niveles kínder a grado 8, grado 9 a grado 12, y universidad, de 2000 a 2006. Graficar estos datos usando gráficas de línea, de barras y de columna apilada.

1.69 Graficar los datos de la tabla 1.12 en una gráfica de columnas 100% apiladas.

Tabla 1.12

Año	2000	2001	2002	2003	2004	2005	2006
Kinder a grado 8	33 852	34 029	34 098	34 065	33 882	33 680	33 507
Grados 9 a 12	13 804	13 862	14 004	14 169	14 483	14 818	15 021
Universidad	12 091	12 225	12 319	12 420	12 531	12 646	12 768

Fuente: U.S. National Center for Educational Statistics and Projections of Education Statistic, annual.

1.70 En la tabla 1.13 se muestra el estado civil de varones y mujeres (18 años o mayores) en Estados Unidos en 1995. Graficar estos datos en: *a*) gráficas de pastel de un mismo diámetro y *b*) una gráfica a elegir.

Tabla 1.13

Estado civil	Varones (porcentaje del total)	Mujeres (porcentaje del total)
Solteros	26.8	19.4
Casados	62.7	59.2
Viudos	2.5	11.1
Divorciados	8.0	10.3

Fuente: U.S. Bureau of Census–Current Population Reports.

1.71 En la tabla 1.14 se da la cantidad de reclusos menores de 18 años en las prisiones estatales de Estados Unidos, de 2001 a 2005. Graficar estos datos en el tipo adecuado de gráficas.

Tabla 1.14

Año	2001	2002	2003	2004	2005
Cantidad	3 147	3 038	2 741	2 485	2 266

1.72 En la tabla 1.15 se da la cantidad (en millones) de visitas al Insituto Smithsoniano, del 2001 al 2005. Con estos datos, construir una gráfica de barras.

Tabla 1.15

Año	2001	2002	2003	2004	2005
Cantidad	32	26	24	20	24

1.73 En la tabla 1.16 se presentan las poblaciones de los siete países más poblados del mundo en 1997. Con estos datos, elaborar una gráfica de pastel.

Tabla 1.16

País	China	India	Estados Unidos	Indonesia	Brasil	Rusia	Pakistán
Población (millones)	1 222	968	268	210	165	148	132

Fuente: U.S. Bureau of the Census, International database.

1.74 Un *diagrama de Pareto* es una gráfica de barras ordenadas de mayor a menor, de izquierda a derecha. Con los datos de la tabla 1.16, construir un diagrama de Pareto.

1.75 En la tabla 1.17 se dan las áreas, en millones de millas cuadradas, de los océanos del mundo. Graficar estos datos en una gráfica: *a*) de barras, *b*) de pastel.

Tabla 1.17

Océano	Pacífico	Atlántico	Índico	Antártico	Ártico
Área (millones de millas cuadradas)	63.8	31.5	28.4	7.6	4.8

Fuente: Naciones Unidas.

ECUACIONES

1.76 Resolver las ecuaciones siguientes:

a) $16 - 5c = 36$

b) $2Y - 6 = 4 - 3Y$

c) $4(X - 3) - 11 = 15 - 2(X + 4)$

d) $3(2U + 1) = 5(3 - U) + 3(U - 2)$

e) $3[2(X + 1) - 4] = 10 - 5(4 - 2X)$

f) $\frac{2}{5}(12 + Y) = 6 - \frac{1}{4}(9 - Y)$

1.77 Resolver las siguientes ecuaciones simultáneas:

a)
$$2a + b = 10$$
$$7a - 3b = 9$$

b)
$$3a + 5b = 24$$
$$2a + 3b = 14$$

c)
$$8X - 3Y = 2$$
$$3X + 7Y = -9$$

d)
$$5A - 9B = -10$$
$$3A - 4B = 16$$

e)
$$2a + b - c = 2$$
$$3a - 4b + 2c = 4$$
$$4a + 3b - 5c = -8$$

f)
$$5X + 2Y + 3Z = -5$$
$$2X - 3Y - 6Z = 1$$
$$X + 5Y - 4Z = 22$$

g)
$$3U - 5V + 6W = 7$$
$$5U + 3V - 2W = -1$$
$$4U - 8V + 10W = 11$$

1.78 *a*) Graficar las ecuaciones $5X + 2Y = 4$ y $7X - 3Y = 23$ en el mismo conjunto de ejes coordenados.

b) A partir de la gráfica, determinar la solución de estas dos ecuaciones simultáneas.

c) Usar los procedimientos de los incisos *a*) y *b*) para obtener las soluciones de las ecuaciones simultáneas *a*) a *d*) del problema 1.77.

1.79 *a*) Usar la gráfica del problema 1.61*a*) para resolver la ecuación $2X^2 + X - 10 = 0$. (*Sugerencia:* Encontrar los valores de X en los que la parábola cruza el eje X: es decir, en los que Y vale 0.)

b) Emplear el método del inciso *a*) para resolver $3X^2 - 4X - 5 = 0$.

1.80 Las soluciones de una ecuación cuadrática $aX^2 + bX + c = 0$ se obtienen mediante la *fórmula cuadrática*:

$$X = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Empleando esta fórmula, encontrar las soluciones de: *a*) $3X^2 - 4X - 5 = 0$, *b*) $2X^2 - X - 10 = 0$, *c*) $5X^2 + 10X = 7$ y *d*) $X^2 + 8X + 25 = 0$.

DESIGUALDADES

1.81 Utilizando los símbolos de desigualdad, ordenar los números -4.3, -6.15, 2.37, 1.52 y -1.5 en: *a*) en orden creciente, *b*) en orden decreciente de magnitud.

1.82 Usar los símbolos de desigualdad para expresar cada una de las afirmaciones siguientes.

a) El número N de niños está entre 30 y 50 inclusive.

b) El número S de puntos en un par de dados no es menor a 7.

c) X es mayor o igual a -4 y menor que 3.

d) P vale a lo mucho 5.

e) X es mayor que Y aumentada en 2.

1.83 Resolver cada una de las desigualdades siguientes:

a) $3X \geq 12$

b) $4X < 5X - 3$

c) $2N + 15 > 10 + 3N$

d) $3 + 5(Y - 2) \leq 7 - 3(4 - Y)$

e) $-3 \leq \frac{1}{5}(2X + 1) \leq 3$

f) $0 < \frac{1}{2}(15 - 5N) \leq 12$

g) $-2 \leq 3 + \frac{1}{2}(a - 12) < 8$

LOGARITMOS Y PROPIEDADES DE LOS LOGARITMOS

1.84 Encontrar los logaritmos comunes:

a) log(10) *b*) log(100) *c*) log(1 000) *d*) log(0.1) *e*) log(0.01)

1.85 Encontrar los logaritmos naturales de los siguientes números a cuatro lugares decimales:

a) ln(e) *b*) ln(10) *c*) ln(100) *d*) ln(1 000) *e*) ln(0.1)

1.86 Encontrar los logaritmos:

a) $\log_4 4$ *b*) $\log_5 25$ *c*) $\log_6 216$ *d*) $\log_7 2\,401$ *e*) $\log_8 32\,768$

1.87 Usar EXCEL para hallar los logaritmos siguientes. Dar la respuesta y los comandos.

a) $\log_4 5$ *b*) $\log_5 24$ *c*) $\log_6 215$ *d*) $\log_7 8$ *e*) $\log_8 9$

1.88 Repetir el problema 1.87 usando MAPLE. Dar la respuesta y los comandos de MAPLE.

1.89 Emplear las propiedades de los logaritmos para escribir la expresión siguiente en forma de sumas y diferencias de logaritmos: $\ln((a^3b^4)/c^5)$

1.90 Emplear las propiedades de los logaritmos para escribir la expresión siguiente en forma de sumas y diferencias de logaritmos: $\log((xyz)/w^3)$

1.91 Transformar la siguiente expresión en una expresión que contenga un solo logaritmo: $5\ln(a) - 4\ln(b) + \ln(c) + \ln(d)$

1.92 Transformar la siguiente expresión en una expresión que contenga un solo logaritmo: $\log(u) + \log(v) + \log(w) - 2\log(x) - 3\log(y) - 4\log(z)$.

ECUACIONES LOGARÍTMICAS

1.93 Encontrar la solución de $\log(3x - 4) = 2$

1.94 Encontrar la solución de $\ln(3x^2 - x) = \ln(10)$

1.95 Encontrar la solución de $\log(w - 2) - \log(2w + 7) = \log(w + 2)$

1.96 Encontrar la solución de $\ln(3x + 5) + \ln(2x - 5) = 12$

1.97 Usar MAPLE o EXCEL para encontrar la solución de $\ln(2x) + \log(3x - 1) = 10$

1.98 Usar MAPLE o EXCEL para encontrar la solución de $\log(2x) + \ln(3x - 1) = 10$

1.99 Usar MAPLE o EXCEL para encontrar la solución de $\ln(3x) - \log(x) = \log_2 3$

1.100 Usar MAPLE o EXCEL para encontrar la solución de $\log_2(3x) - \log(x) = \ln(3)$

DISTRIBUCIONES DE FRECUENCIAS

2

DATOS EN BRUTO

Los *datos en bruto* son los datos recolectados que aún no se han organizado. Por ejemplo, las estaturas de 100 estudiantes tomados de la lista alfabética de una universidad.

ORDENACIONES

Ordenación se le llama a los datos numéricos en bruto dispuestos en orden creciente o decreciente de magnitud. A la diferencia entre el número mayor y el número menor se le conoce como el *rango* de los datos. Por ejemplo, si la estatura mayor en los 100 estudiantes es 74 pulgadas (in) y la menor es 60 in, el rango es $74 - 60 = 14$ pulgadas (in).

DISTRIBUCIONES DE FRECUENCIA

Al organizar una gran cantidad de datos en bruto, suele resultar útil distribuirlos en *clases* o *categorías* y determinar la cantidad de datos que pertenece a cada clase; esta cantidad se conoce como la *frecuencia de clase*. A la disposición tabular de los datos en clases con sus respectivas frecuencias de clase se le conoce como *distribución de frecuencias* o *tabla de frecuencias*. La tabla 2.1 es una distribución de frecuencias de las estaturas (registradas a la pulgada más cercana) de 100 estudiantes de la universidad XYZ.

Tabla 2.1 Estaturas de 100 estudiantes de la universidad XYZ

Estatura (in)	Cantidad de estudiantes
60-62	5
63-65	18
66-68	42
69-71	27
72-74	8
	Total 100

La primera clase (o categoría), por ejemplo, consta de las estaturas que van desde 60 hasta 62 pulgadas y queda identificada por el símbolo 60-62. Como hay cinco estudiantes cuyas estaturas pertenecen a esta clase, la frecuencia de clase correspondiente es 5.

A los datos organizados y resumidos como en la distribución de frecuencias anterior se les llama *datos agrupados.* Aunque al agrupar los datos se pierden muchos de los detalles originales de los datos, esto tiene la ventaja de que se obtiene una visión general clara y se hacen evidentes las relaciones.

INTERVALOS DE CLASE Y LÍMITES DE CLASE

Al símbolo que representa una clase, como 60-62 en la tabla 2.1, se le conoce como *intervalo de clase.* A los números de los extremos, 60 y 62, se les conoce como *límites de clase;* el número menor (60) es el *límite inferior de clase*, y el número mayor (62) es el *límite superior de clase.* Los términos *clase* e *intervalo de clase* se suelen usar indistintamente, aunque el intervalo de clase en realidad es un símbolo para la clase.

Un intervalo de clase que, por lo menos teóricamente, no tenga indicado el límite de clase superior o el límite de clase inferior, se conoce como *intervalo de clase abierto.* Por ejemplo, al considerar grupos de edades de personas, un intervalo que sea "65 años o mayores" es un intervalo de clase abierto.

FRONTERAS DE CLASE

Si las estaturas se registran a la pulgada más cercana, el intervalo de clase 60-62 comprende teóricamente todas las mediciones desde 59.5000 hasta 62.5000 in. Estos números que se indican brevemente mediante los números exactos 59.5 y 62.5 son las *fronteras de clase* o los *límites de clase reales;* el menor de los números (59.5) es la *frontera inferior de clase* y el número mayor (62.5) es la *frontera superior de clase.*

En la práctica, las fronteras de clase se obtienen sumando el límite superior de un intervalo de clase al límite inferior del intervalo de clase inmediato superior y dividiendo entre 2.

Algunas veces, las fronteras de clase se usan para representar a las clases. Por ejemplo, las clases de la tabla 2.1 pueden indicarse como 59.5-62.5, 62.5-65.5, etc. Para evitar ambigüedades cuando se usa esta notación, las fronteras de clase no deben coincidir con las observaciones. Por lo tanto, si una observación es 62.5, no es posible decidir si pertenece al intervalo 59.5-62.5 o al intervalo 62.5-65.5

TAMAÑO O AMPLITUD DE UN INTERVALO DE CLASE

El tamaño, o la amplitud, de un intervalo de clase es la diferencia entre sus fronteras superior e inferior y se le conoce también como *amplitud de clase, tamaño de clase* o *longitud de clase.* Si en una distribución de frecuencia todos los intervalos de clase tienen la misma amplitud, esta amplitud común se denota c. En este caso, c es igual a la diferencia entre dos límites inferiores de clases sucesivas o entre dos límites superiores de clases sucesivas. Por ejemplo, en los datos de la tabla 2.1, el intervalo de clase es $c = 62.5 - 59.5 = 65.5 - 62.5 = 3$.

LA MARCA DE CLASE

La *marca de clase* es el punto medio del intervalo de clase y se obtiene sumando los límites de clase inferior y superior y dividiendo entre 2. Así, la marca de clase del intervalo 60-62 es $(60 + 62)/2 = 61$. A la marca de clase también se le conoce como *punto medio de clase.*

Para los análisis matemáticos posteriores, se supone que todas las observaciones que pertenecen a un intervalo de clase dado coinciden con la marca de clase. Así, se considera que todas las estaturas en el intervalo de clase 60-62 in son de 61 in.

REGLAS GENERALES PARA FORMAR UNA DISTRIBUCIÓN DE FRECUENCIAS

1. En el conjunto de los datos en bruto, se determina el número mayor y el número menor y se halla, así, el rango (la diferencia entre los números mayor y menor).

2. Se divide el rango en una cantidad adecuada de intervalos de clase de una misma amplitud. Si esto no es posible, se usan intervalos de clase de diferentes amplitudes o intervalos de clase abiertos (ver problema 2.12). La cantidad de intervalos suele ser de 5 a 20, dependiendo de los datos. Los intervalos de clase también suelen elegirse de manera que las marcas de clase (o puntos medios de clase) coincidan con datos observados. Esto tiende a disminuir el llamado *error de agrupamiento* en los análisis matemáticos subsiguientes. En cambio, las fronteras de clase no deben coincidir con datos observados.
3. Se determina la cantidad de observaciones que caen dentro de cada intervalo de clase; es decir, se encuentran las frecuencias de clase. La mejor manera de hacer esto es utilizando una *hoja de conteo* (ver problema 2.8).

HISTOGRAMAS Y POLÍGONOS DE FRECUENCIAS

Los histogramas y los polígonos de frecuencias son dos representaciones gráficas de las distribuciones de frecuencias.

1. Un *histograma* o *histograma de frecuencias* consiste en un conjunto de rectángulos que tienen: *a*) sus bases sobre un eje horizontal (el eje X), con sus centros coincidiendo con las marcas de clase de longitudes iguales a la amplitud del intervalo de clase, y *b*) áreas proporcionales a las frecuencias de clase.
2. Un *polígono de frecuencias* es una gráfica de línea que presenta las frecuencias de clase graficadas contra las marcas de clase. Se puede obtener conectando los puntos medios de las partes superiores de los rectángulos de un histograma.

En las figuras 2.1 y 2.2 se muestran el histograma y el polígono de frecuencias correspondientes a la distribución de frecuencias de las estaturas presentada en la tabla 2.1.

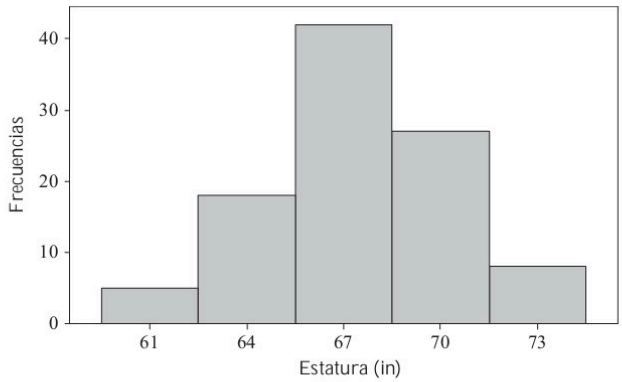


Figura 2-1 MINITAB, histograma que muestra los puntos medios y las frecuencias de clase.

Obsérvese en la figura 2.2 cómo el polígono de frecuencias se ha anclado por sus extremos, es decir, en 58 y 76.

DISTRIBUCIONES DE FRECUENCIAS RELATIVAS

La *frecuencia relativa* de una clase es la frecuencia de la clase dividida entre la suma de las frecuencias de todas las clases y generalmente se expresa como porcentaje. Por ejemplo, en la tabla 2.1, la frecuencia relativa de la clase 66-68 es $42/100 = 42\%$. Por supuesto, la suma de las frecuencias relativas de todas las clases es 1, o 100%.

Si en la tabla 2.1 las frecuencias se sustituyen por frecuencias relativas, la tabla que se obtiene es una *distribución de frecuencias relativas*, *distribución porcentual* o *tabla de frecuencias relativas*.

Las representaciones gráficas de las distribuciones de frecuencias relativas se obtienen a partir de los histogramas o polígonos de frecuencias, cambiando únicamente, en la escala vertical, las frecuencias por las frecuencias relativas y conservando la gráfica exactamente igual. A las gráficas que se obtienen se les llama *histogramas de frecuencias relativas* (o *histogramas porcentuales*) y *polígonos de frecuencias relativas* (o *polígonos porcentuales*), respectivamente.

DISTRIBUCIONES DE FRECUENCIAS ACUMULADAS Y OJIVAS

A la suma de todas las frecuencias menores que la frontera superior de un intervalo de clase dado se le llama *frecuencia acumulada* hasta ese intervalo de clase inclusive. Por ejemplo, en la tabla 2.1, la frecuencia acumulada hasta el intervalo de clase 66-68 inclusive es $5 + 18 + 42 = 65$, lo que significa que 65 estudiantes tienen una estatura menor a 68.5 in.

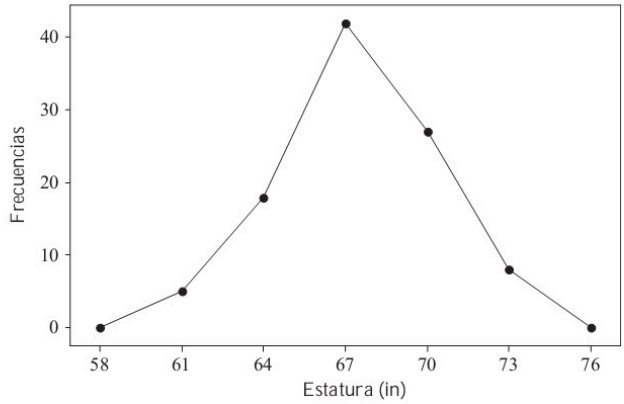


Figura 2-2 MINITAB, polígono de frecuencias de las estaturas de los estudiantes.

A una tabla en la que se presentan las frecuencias acumuladas se le llama *distribución de frecuencias acumuladas*, *tabla de frecuencias acumuladas* o simplemente *distribución acumulada*, y se presenta en la tabla 2.2 para la distribución de las estaturas de los estudiantes de la tabla 2.1.

Tabla 2.2

Estatura (in)	Cantidad de estudiantes
Menos de 59.5	0
Menos de 62.5	5
Menos de 65.5	23
Menos de 68.5	65
Menos de 71.5	92
Menos de 74.5	100

Una gráfica que muestra las frecuencias acumuladas menores de cada frontera superior de clase respecto a cada frontera superior de clase se le conoce como *gráfica de frecuencias acumuladas* u *ojiva*. En algunas ocasiones se desea considerar distribuciones de frecuencias mayores o iguales que la frontera inferior de cada intervalo de clase. Como en ese caso se consideran las estaturas de 59.5 in o más, de 62.5 in o más, etc., a estas distribuciones se les suele llamar *distribuciones acumuladas "o más que"*, en tanto que las distribuciones consideradas antes son *distribuciones acumuladas "o menos que"*. Una puede obtenerse fácilmente de la otra. A las ojivas correspondientes se les llama ojivas "más que" y ojivas "menos que". Aquí, siempre que se hable de distribuciones acumuladas o de ojivas, sin más, se tratará del tipo "menos que".

DISTRIBUCIONES DE FRECUENCIAS ACUMULADAS RELATIVAS Y OJIVAS PORCENTUALES

La *frecuencia acumulada relativa* o *frecuencia acumulada porcentual* es la frecuencia acumulada dividida entre la suma de todas las frecuencias (frecuencia total). Por ejemplo, la frecuencia acumulada relativa de las estaturas menores que 68.5 in es $65/100 = 0.65$ o 65%, lo que significa que 65% de los estudiantes tienen estaturas menores a

68.5 in. Si en la tabla 2.2 se emplean las frecuencias acumuladas relativas en lugar de las frecuencias acumuladas, se obtiene una *distribución de frecuencias acumuladas relativas* (o *distribución acumulada porcentual*) y una *gráfica de frecuencias acumuladas relativas* (u *ojiva porcentual*), respectivamente.

CURVAS DE FRECUENCIAS Y OJIVAS SUAVIZADAS

Suele considerarse que los datos recolectados pertenecen a una muestra obtenida de una población grande. Como de esta población se pueden obtener muchas observaciones, teóricamente es posible (si son datos continuos) elegir intervalos de clase muy pequeños y, a pesar de eso, tener un número adecuado de observaciones que caigan en cada clase. De esta manera, cuando se tienen poblaciones grandes puede esperarse que los polígonos de frecuencias, o los polígonos de frecuencias relativas, correspondientes a estas poblaciones estén formados por una gran cantidad de pequeños segmentos de recta de manera que sus formas se aproximen a las de unas curvas, a las cuales se les llama *curvas de frecuencias* o *curvas de frecuencias relativas,* respectivamente.

Es razonable esperar que estas curvas teóricas puedan ser aproximadas suavizando los polígonos de frecuencias o los polígonos de frecuencias relativas de la muestra; esta aproximación mejorará a medida que aumenta el tamaño de la muestra. Ésta es la razón por la que a las curvas de frecuencias se les suele llamar polígonos de *frecuencias suavizados*.

De igual manera, suavizando las gráficas de frecuencias acumuladas u ojivas, se obtienen *ojivas suavizadas.* Por lo general, es más fácil suavizar una ojiva que un polígono de frecuencias.

TIPOS DE CURVAS DE FRECUENCIAS

Las curvas de frecuencias que surgen en la práctica toman ciertas formas características, como las que se muestran en la figura 2-3.

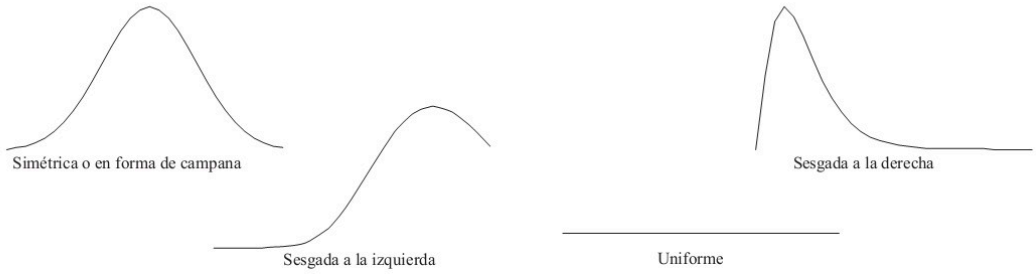


Figura 2-3 Cuatro distribuciones que se encuentran con por lo común.

1. Las curvas *simétricas* o *en forma de campana* se caracterizan porque las observaciones equidistantes del máximo central tienen la misma frecuencia. Las estaturas tanto de hombres como de mujeres adultos tienen distribuciones en forma de campana.
2. Las curvas que tienen colas hacia la izquierda se dice que son *sesgadas a la izquierda*. Las curvas de la cantidad de años que viven hombres y mujeres son sesgadas a la izquierda. Pocos mueren jóvenes y la mayoría muere entre los 60 y los 80 años. En general, las mujeres viven en promedio diez años más que los hombres.
3. Las curvas que tienen colas hacia la derecha se dice que son *sesgadas a la derecha*. Las curvas de las edades a las que se casan tanto hombres como mujeres son sesgadas a la derecha. La mayoría se casa entre los veinte y treinta años y pocos se casan alrededor de cuarenta, cincuenta, sesenta o setenta años.
4. Las curvas que tienen aproximadamente las mismas frecuencias para todos sus valores se dice que son curvas *distribuidas uniformemente*. Por ejemplo, las máquinas dispensadoras de refresco lo hacen de manera uniforme entre 15.9 y 16.1 onzas.
5. Las curvas de frecuencias en *forma de J* o en *forma de J inversa* son curvas en las que el máximo se presenta en uno de sus extremos.
6. Las curvas de frecuencias en *forma de U* son curvas que tienen un máximo en cada extremo y un mínimo en medio.
7. Las curvas *bimodales* son curvas que tienen dos máximos.
8. Las curvas *multimodales* tienen más de dos máximos.

PROBLEMAS RESUELTOS

ORDENACIONES

2.1 *a*) Disponer los números 17, 45, 38, 27, 6, 48, 11, 57, 34 y 22 en una ordenación.
b) Determinar el rango de estos números.

SOLUCIÓN

a) En orden ascendente de magnitud, la ordenación es: 6, 11, 17, 22, 27, 34, 38, 45, 48, 57. En orden descendente de magnitud, la ordenación es: 57, 48, 45, 38, 34, 27, 22, 17, 11, 6.
b) Como el número mayor es 57 y el número menor es 6, el rango es $57 - 6 = 51$.

2.2 En la tabla siguiente se presentan las calificaciones finales que obtuvieron en matemática 80 alumnos de una universidad.

68	84	75	82	68	90	62	88	76	93
73	79	88	73	60	93	71	59	85	75
61	65	75	87	74	62	95	78	63	72
66	78	82	75	94	77	69	74	68	60
96	78	89	61	75	95	60	79	83	71
79	62	67	97	78	85	76	65	71	75
65	80	73	57	88	78	62	76	53	74
86	67	73	81	72	63	76	75	85	77

De acuerdo con esta tabla, encontrar:

a) La calificación más alta.
b) La calificación más baja.
c) El rango.
d) Las calificaciones de los cinco mejores estudiantes.
e) Las calificaciones de los cinco peores estudiantes.
f) La calificación del alumno que tiene el décimo lugar entre las mejores calificaciones.
g) El número de estudiantes que obtuvieron 75 o más.
h) El número de estudiantes que obtuvieron 85 o menos.
i) El porcentaje de los estudiantes que obtuvieron calificaciones mayores a 65 pero no mayores a 85.
j) Las calificaciones que no aparecen en esta tabla.

SOLUCIÓN

Como algunas de estas preguntas son tan minuciosas, es mejor construir primero una ordenación. Esto se hace dividiendo los datos, de manera adecuada, en clases y colocando cada número de la tabla en su clase correspondiente, como se ve en la tabla 2.3, llamada *tabla de entradas*. Después, los números de cada clase se disponen en una ordenación, como se muestra en la tabla 2.4, con lo que se obtiene la ordenación deseada. Consultando la tabla 2.4 es relativamente fácil responder a las preguntas anteriores.

a) La calificación más alta es 97.
b) La calificación más baja es 53.
c) El rango es $97 - 53 = 44$
d) Las calificaciones de los cinco mejores estudiantes son 97, 96, 95, 95 y 94.

Tabla 2.3

50-54	53
55-59	59, 57
60-64	62, 60, 61, 62, 63, 60, 61, 60, 62, 62, 63
65-69	68, 68, 65, 66, 69, 68, 67, 65, 65, 67
70-74	73, 73, 71, 74, 72, 74, 71, 71, 73, 74, 73, 72
75-79	75, 76, 79, 75, 75, 78, 78, 75, 77, 78, 75, 79, 79, 78, 76, 75, 78, 76, 76, 75, 77
80-84	84, 82, 82, 83, 80, 81
85-89	88, 88, 85, 87, 89, 85, 88, 86, 85
90-94	90, 93, 93, 94
95-99	95, 96, 95, 97

Tabla 2.4

50-54	53
55-59	57, 59
60-64	60, 60, 60, 61, 61, 62, 62, 62, 62, 63, 63
65-69	65, 65, 65, 66, 67, 67, 68, 68, 68, 69
70-74	71, 71, 71, 72, 72, 73, 73, 73, 73, 74, 74, 74
75-79	75, 75, 75, 75, 75, 75, 75, 76, 76, 76, 76, 77, 77, 78, 78, 78, 78, 78, 79, 79, 79
80-84	80, 81, 82, 82, 83, 84
85-89	85, 85, 85, 86, 87, 88, 88, 88, 89
90-94	90, 93, 93, 94
95-99	95, 95, 96, 97

e) Las calificaciones de los cinco peores estudiantes son 53, 57, 59, 60 y 60.

f) La calificación del alumno que tiene el décimo lugar entre las mejores calificaciones es 88.

g) La cantidad de estudiantes que obtuvieron 75 o más es 44.

h) La cantidad de estudiantes que obtuvieron menos de 85 es 63.

i) El porcentaje de estudiantes que obtuvieron calificaciones mayores a 65 pero no mayores a 85 es $49/80 = 61.2\%$.

j) Las calificaciones que no aparecen en esta tabla son desde 0 hasta 52, 54, 55, 56, 58, 64, 70, 91, 92, 98, 99 y 100.

DISTRIBUCIONES DE FRECUENCIAS, HISTOGRAMAS Y POLÍGONOS DE FRECUENCIAS

2.3 La tabla 2.5 muestra una distribución de frecuencias de los salarios semanales de 65 empleados de la empresa P&R. Con los datos de esta tabla, determinar:

a) El límite inferior de la sexta clase.

b) El límite superior de la cuarta clase.

c) La marca de clase (o punto medio de clase) de la tercera clase.

d) Las fronteras de clase de la quinta clase.

e) La amplitud del intervalo de la quinta clase.

f) La frecuencia de la tercera clase.

g) La frecuencia relativa de la tercera clase.

h) El intervalo de clase de mayor frecuencia. A este intervalo se le suele llamar *intervalo de clase modal* y a su frecuencia se le conoce como *frecuencia de la clase modal*.

Tabla 2.5

Salarios	Número de empleados
\$250.00-\$259.99	8
\$260.00-\$269.99	10
\$270.00-\$279.99	16
\$280.00-\$289.99	14
\$290.00-\$299.99	10
\$300.00-\$309.99	5
\$310.00-\$319.99	2
	Total 65

i) El porcentaje de empleados que gana menos de \$280.00 por semana.

j) El porcentaje de empleados que gana menos de \$300.00 por semana, pero por lo menos \$260.00 por semana.

SOLUCIÓN

a) \$300.00.

b) \$289.99.

c) La marca de clase (o punto medio de clase) de la tercera clase $= \frac{1}{2}(\$270.00 + \$279.99) = \$274.995$. Para propósitos prácticos, esta cantidad se redondea a \$275.00.

d) La frontera inferior de la quinta clase $= \frac{1}{2}(\$290.00 + \$289.99) = \$289.995$. La frontera superior de la quinta clase $= \frac{1}{2}(\$299.99 + \$300.00) = \$299.995$.

e) La amplitud del intervalo de la quinta clase = frontera superior de la quinta clase – frontera inferior de la quinta clase = \$299.995 − \$289.985 = \$10.00. En este caso, todos los intervalos de clase son del mismo tamaño: \$10.00.

f) 16.

g) $16/65 = 0.246 = 24.6\%$.

h) \$270.00 − \$279.99.

i) El número total de empleados que gana menos de \$280 por semana $= 16 + 10 + 8 = 34$. El porcentaje de empleados que gana menos de \$280 por semana $= 34/65 = 52.3\%$.

j) El número de empleados que gana menos de \$300 por semana pero más de \$260 por semana $= 10 + 14 + 16 + 10 = 50$. El porcentaje de empleados que gana menos de \$300 por semana, pero por lo menos \$260 por semana $= 50/65 = 76.9\%$.

2.4 Si las marcas de clase en una distribución de frecuencias de pesos de estudiantes son 128, 137, 146, 155, 164, 173 y 182 libras, encuentre: *a*) la amplitud del intervalo de clase, *b*) las fronteras de clase y *c*) los límites de clase, suponiendo que los pesos se hayan redondeado a la libra más cercana.

SOLUCIÓN

a) La amplitud del intervalo de clase = diferencia entre marcas sucesivas de clase = 137 − 128 = 146 − 137 = etc. = 9 lb.

b) Como todos los intervalos de clase tienen la misma amplitud, las fronteras de clase están a medio camino entre dos marcas de clase y por lo tanto se tienen los valores

$$\tfrac{1}{2}(128 + 137), \tfrac{1}{2}(137 + 146), \ldots, \tfrac{1}{2}(173 + 182) \qquad \text{o bien} \qquad 132.5, 141.5, 150.5, \ldots, 177{,}5 \text{ lb}$$

La frontera de la primera clase es $132.5 - 9 = 123.5$ y la frontera de la última clase es $177.5 + 9 = 186.5$, ya que la amplitud común de los intervalos de clase es 9 lb. Por lo tanto, todas las fronteras de clase son:

$$123.5,\ 132.5,\ 141.5,\ 150.5,\ 159.5,\ 168.5,\ 177.5,\ 186.5 \text{ lb}$$

c) Como los límites de clase son enteros, se eligen éstos como los enteros más cercanos a las fronteras de clase, es decir, 123, 124, 132, 133, 141, 142... Así, los límites de la primera clase son 124-132, de la siguiente, 133-141, etcétera.

2.5 Se toma una muestra de la cantidad de tiempo, en horas por semana, que los estudiantes universitarios usan su celular. Usando SPSS, la secuencia **"Analyze ⇒ Descripive Statistics ⇒ Frequencies"** da el resultado mostrado en la figura 2-4.

Tiempo

		Frecuencias	Porcentajes	Porcentajes válidos	Porcentajes acumulados
Válido	3.00	3	6.0	6.0	6.0
	4.00	3	6.0	6.0	12.0
	5.00	5	10.0	10.0	22.0
	6.00	3	6.0	6.0	28.0
	7.00	4	8.0	8.0	36.0
	8.00	4	8.0	8.0	44.0
	9.00	3	6.0	6.0	50.0
	10.00	4	8.0	8.0	58.0
	11.00	2	4.0	4.0	62.0
	12.00	2	4.0	4.0	66.0
	13.00	3	6.0	6.0	72.0
	14.00	1	2.0	2.0	74.0
	15.00	2	4.0	4.0	78.0
	16.00	5	10.0	10.0	88.0
	17.00	2	4.0	4.0	92.0
	18.00	1	2.0	2.0	94.0
	19.00	2	4.0	4.0	98.0
	20.00	1	2.0	2.0	100.0
	Total	50	100.0	100.0	

Figura 2-4 SPSS, resultados para el problema 2.5.

a) ¿Qué porcentaje usa su celular 15 o menos horas por semana?
b) ¿Qué porcentaje usa su celular 10 o más horas por semana?

SOLUCIÓN

a) El porcentaje acumulado correspondiente a 15 horas es 78%. Es decir, 78% usa su celular 15 horas o menos por semana.

b) El porcentaje acumulado correspondiente a 10 horas es 58%. Es decir, 58% usa su celular 10 horas o menos por semana. Por lo tanto, 42% usa su celular más de 10 horas por semana.

2.6 De 150 mediciones, la menor es 5.18 in y la mayor es 7.44 in. Determinar un conjunto adecuado: *a*) de intervalos de clase, *b*) de fronteras de clase y *c*) de marcas de clase que se pueda usar para elaborar una distribución de frecuencias con estas mediciones.

SOLUCIÓN

El rango es $7.44 - 5.18 = 226$ in. Para un mínimo de cinco intervalos de clases, la amplitud del intervalo de clase es $2.26/5 = 0.45$, aproximadamente, y para un máximo de 20 intervalos de clase, la amplitud del intervalo de clase es $2.26/20 = 0.11$, aproximadamente. Las amplitudes adecuadas para el intervalo de clase, entre 0.11 y 0.45, podrían ser 0.20, 0.30 o 0.40.

a) En las columnas I, II y III de la tabla siguiente se presentan intervalos de clase de amplitud 0.20, 0.30 y 0.40, respectivamente.

I	II	III
5.10-5.29	5.10-5.39	5.10-5.49
5.30-5.49	5.40-5.69	5.50-5.89
5.50-5.69	5.70-5.99	5.90-6.29
5.70-5.89	6.00-6.29	6.30-6.69
5.90-6.09	6.30-6.59	6.70-7.09
6.10-6.29	6.60-6.89	7.10-7.49
6.30-6.49	6.90-7.19	
6.50-6.69	7.20-7.49	
6.70-6.89		
6.90-7.09		
7.10-7.29		
7.30-7.49		

Obsérvese que el límite inferior de cada una de las primeras clases puede ser también distinto a 5.10. Por ejemplo, si en la columna I se empieza con 5.15 como límite inferior, el primer intervalo de clase será 5.15-5.34.

b) Las fronteras de clase correspondientes a las columnas I, II y III del inciso *a*) son:

I 5.095-5.295, 5.295-5.495, 5.495-5.695, . . . , 7.295-7.495

II 5.095-5.395, 5.395-5.695, 5.695-5.995, . . . , 7.195-7.495

III 5.095-5.495, 5.495-5.895, 5.895-6.295, . . . , 7.095-7.495

Obsérvese que estas fronteras de clase son adecuadas, ya que no coinciden con las mediciones observadas.

c) A continuación se dan las marcas de clase correspondientes a las columnas I, II y III del inciso *a*).

I 5.195, 5.395, . . . , 7.395 II 5.245, 5.545, . . . , 7.345 III 5.295, 5.695, . . . , 7.295

Estas marcas de clase tienen la desventaja de no coincidir con mediciones observadas

2.7 Al resolver el problema 2.6*a*), un estudiante elige como intervalos de clase 5.10-5.40, 5.40-5.70, . . . , 6.90-7.20 y 7.20-7.50. ¿Hay algún problema con esta elección?

SOLUCIÓN

Estos intervalos de clase se traslapan en 5.40, 5.70, . . . , 7.20. De esta manera, una medición que se registre, por ejemplo como 5.40, podrá colocarse en cualquiera de los dos primeros intervalos de clases. Algunos justifican esto acordando colocar la mitad de los casos ambiguos en una de las clases y la otra mitad en la otra.

Esta ambigüedad se elimina escribiendo los intervalos de clase como de 5.10 hasta menos de 5.40, de 5.40 hasta menos de 5.70, etc. En este caso, los límites de clase coinciden con las fronteras de clase y las marcas de clase pueden coincidir con datos observados.

En general, siempre que sea posible, se desea evitar que los intervalos de clase se superpongan y escogerlos de manera que las fronteras de clase sean valores que no coincidan con datos observados. Por ejemplo, los intervalos de clase del problema 2.6 pueden ser 5.095-5.395, 5.395-5.695, etc., sin que haya ambigüedad. La desventaja de este caso particular es que las marcas de clase no coincidirán con datos observados.

2.8 En la tabla siguiente se presentan los pesos, dados a la libra más cercana, de 40 estudiantes de una universidad. Elaborar una distribución de frecuencias.

138	164	150	132	144	125	149	157
146	158	140	147	136	148	152	144
168	126	138	176	163	119	154	165
146	173	142	147	135	153	140	135
161	145	135	142	150	156	145	128

SOLUCIÓN

El peso mayor es 176 lb y el peso menor es 119 lb, de manera que el rango es $176 - 119 = 57$ lb. Si se emplean cinco intervalos de clase, la amplitud del intervalo de clase será $57/5 = 11$, aproximadamente; si se usan 20 intervalos de clase, la amplitud de cada intervalo de clase será $57/20 = 3$, aproximadamente.

Una amplitud adecuada para los intervalos de clase es 5 lb. También es conveniente que las marcas de clase sean 120, 125, 130, 135, . . . , lb. Por lo tanto, los intervalos de clase serán 118-122, 123-127, 128-132, . . . Y entonces las fronteras de clase serán 117.5, 122.5, 127.5, . . . , las cuales no coinciden con datos observados.

La distribución de frecuencias buscada se muestra en la tabla 2.6. La columna central, llamada *hoja de conteo*, se usa para tabular las frecuencias de clase a partir de los datos en bruto y suele omitirse en la presentación final de una distribución de frecuencias. No es necesario hacer una ordenación, pero si se cuenta con ella, se puede usar para tabular las frecuencias.

Otro método

Por supuesto, hay otras posibles distribuciones de frecuencias. En la tabla 7.2, por ejemplo, se muestra una distribución de frecuencias que tiene sólo siete clases y en la que el intervalo de clase es de 9 lb.

Tabla 2.6

Peso (lb)	Conteo	Frecuencias
118-122	/	1
123-127	//	2
128-132	//	2
133-137	////	4
138-142	///// /	6
143-147	///// ///	8
148-152	/////	5
153-157	////	4
158-162	//	2
163-167	///	3
168-172	/	1
173-177	//	2
	Total	40

Tabla 2.7

Peso (lb)	Conteo	Frecuencias
118-126	///	3
127-135	/////	5
136-144	///// /////	9
145-153	///// ///// //	12
154-162	/////	5
163-171	////	4
172-180	//	2
	Total	40

2.9 Se toman las estaturas de 45 estudiantes del sexo femenino de una universidad; a continuación se presentan estas estaturas registradas a la pulgada más cercana. Para elaborar un histograma, usar el paquete STATISTIX para estadística.

67	67	64	64	74	61	68	71	69	61	65	64	62	63	59
70	66	66	63	59	64	67	70	65	66	66	56	65	67	69
64	67	68	67	67	65	74	64	62	68	65	65	65	66	67

SOLUCIÓN

Después de ingresar los datos en la hoja de cálculo de STATISTIX, la secuencia **"Statistics ⇒ Summary Statistics ⇒ Histogram"** produce el histograma que se muestra en la figura 2-5.

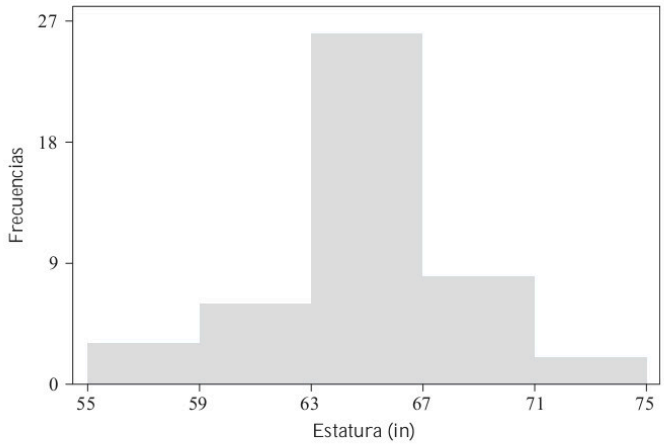


Figura 2-5 STATISTIX, histograma de las estaturas de 45 estudiantes universitarias.

2.10 En la tabla 2.8 se dan las distancias, en millas, que recorren 50 estudiantes del Metropolitan College de su casa a la universidad.

Tabla 2.8 Distancias al Metropolitan College (millas)

4.3	7.0	8.0	3.9	3.7	8.4	2.6	1.0	15.7	3.9
6.5	8.7	0.9	0.9	12.6	4.0	10.3	10.0	6.2	1.1
7.2	8.8	7.8	4.9	2.0	3.0	4.2	3.3	4.8	4.4
7.7	2.4	8.0	8.0	4.6	1.4	2.2	1.9	3.2	4.8
5.0	10.3	12.3	3.8	3.8	6.6	2.0	1.6	4.4	4.3

En la figura 2.6 se muestra el histograma obtenido con SPSS con las distancias de la tabla 2.8. Obsérvese que las clases son 0 a 2, 2 a 4, 4 a 6, 6 a 8, 8 a 10, 10 a 12, 12 a 14 y 14 a 16. Las frecuencias 7, 13, 11, 7, 6, 3, 2 y 1. Un número que cae en el límite inferior de clase se cuenta dentro de esa clase, pero los que caen en el límite superior, se cuentan dentro de la clase siguiente.

a) ¿Cuáles son los valores que pertenecen a la primera clase?
b) ¿Cuáles son los valores que pertenecen a la segunda clase?
c) ¿Cuáles son los valores que pertenecen a la tercera clase?
d) ¿Cuáles son los valores que pertenecen a la cuarta clase?
e) ¿Cuáles son los valores que pertenecen a la quinta clase?
f) ¿Cuáles son los valores que pertenecen a la sexta clase?
g) ¿Cuáles son los valores que pertenecen a la séptima clase?
h) ¿Cuáles son los valores que pertenecen a la octava clase?

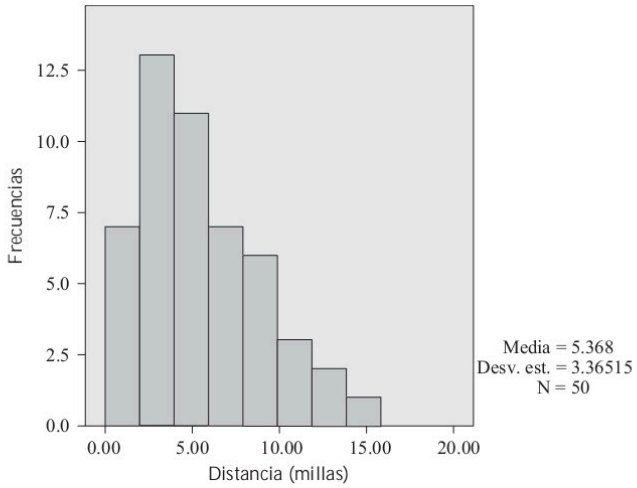


Figura 2-6 SPSS, histograma de las distancias al Metropolitan College.

SOLUCIÓN

a) 0.9, 0.9, 1.0, 1.1, 1.4, 1.6, 1.9
b) 2.0, 2.0, 2.2, 2.4, 2.6, 3.0, 3.2, 3.3, 3.7, 3.8, 3.8, 3.9, 3.9
c) 4.0, 4.2, 4.3, 4.3, 4.4, 4.4, 4.6, 4.8, 4.8, 4.9, 5.0
d) 6.2, 6.5, 6.6, 7.0, 7.2, 7.7, 7.8
e) 8.0, 8.0, 8.0, 8.4, 8.7, 8.8
f) 10.0, 10.3, 10.3
g) 12.3, 12.6
h) 15.7

2.11 En la figura 2-7 se muestra un histograma obtenido con SAS, con las distancias de la tabla 2.8. Se muestran los puntos medios (marcas de clase) de los intervalos de clase. Las clases son 0 a 2.5, 2.5 a 5.0, 5.0 a 7.5, 7.5 a 10.0, 10 a 12.5, 12.5 a 15.0, 15.0 a 17.5, 17.5 a 20.0. Los números que caen en el límite inferior de clase se cuentan dentro de esa clase, pero si caen en el límite superior se cuentan dentro de la clase siguiente.

a) ¿Cuáles son los valores (de la tabla 2.8) que pertenecen a la primera clase?
b) ¿Cuáles son los valores que pertenecen a la segunda clase?
c) ¿Cuáles son los valores que pertenecen a la tercera clase?
d) ¿Cuáles son los valores que pertenecen a la cuarta clase?
e) ¿Cuáles son los valores que pertenecen a la quinta clase?
f) ¿Cuáles son los valores que pertenecen a la sexta clase?
g) ¿Cuáles son los valores que pertenecen a la séptima clase?

SOLUCIÓN

a) 0.9, 0.9, 1.0, 1.1, 1.4, 1.6, 1.9, 2.0, 2.0, 2.2, 2.4
b) 2.6, 3.0, 3.2, 3.3, 3.7, 3.8, 3.8, 3.9, 3.9, 4.0, 4.2, 4.3, 4.3, 4.4, 4.4, 4.6, 4.8, 4.8, 4.9
c) 5.0, 6.2, 6.5, 6.6, 7.0, 7.2
d) 7.7, 7.8, 8.0, 8.0, 8.0, 8.4, 8.7, 8.8

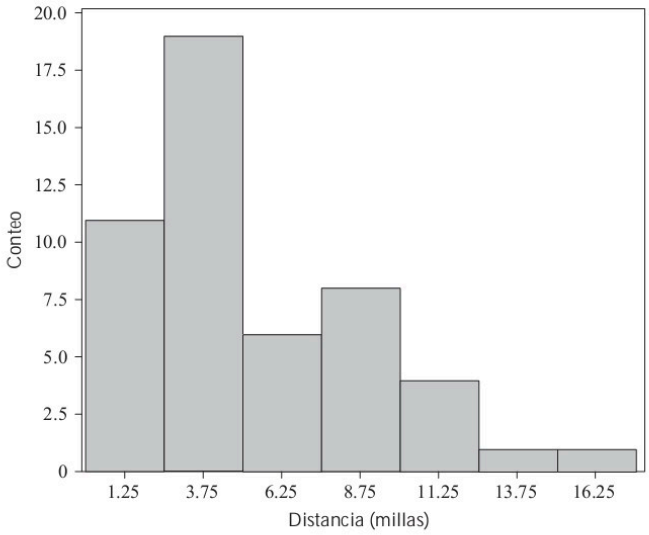


Figura 2-7 SAS, histograma con las distancias al Metropolitan College.

e) 10.0, 10.3, 10.3, 12.3
f) 12.6
g) 15.7

2.12 La empresa P&R (problema 2.3) contrata cinco empleados nuevos, cuyos salarios semanales son $285.34, $316.83, $335.78, $356.21 y $374.50. Construir una distribución de frecuencias con los salarios de los 70 empleados.

SOLUCIÓN

En las tablas 2.9, *a*) y *b*), se presentan varias distribuciones de frecuencias posibles.

Tabla 2.9*a*)

Salarios	Frecuencias
$250.00-$259.99	8
260.00-269.99	10
270.00-279.99	16
280.00-289.99	15
290.00-299.99	10
300.00-309.99	5
310.00-319.99	3
320.00-329.99	0
330.00-339.99	1
340.00-349.99	0
350.00-359.99	1
360.00-369.99	0
370.00-379.99	1
	Total 70

Tabla 2.9*b*)

Salarios	Frecuencias
$250.00-$259.99	8
260.00-269.99	10
270.00-279.99	16
280.00-289.99	15
290.00-299.99	10
300.00-309.99	5
310.00-319.99	3
320.00 y más	3
	Total 70

Tabla 2.9*c*)

Salarios	Frecuencias
$250.00-$269.99	18
270.00-289.99	31
290.00-309.99	15
310.00-329.99	3
330.00-349.99	1
350.00-369.99	1
370.00-389.99	1
	Total 70

Tabla 2.9*d*)

Salarios	Frecuencias
$250.00-$259.99	8
260.00-269.99	10
270.00-279.99	16
280.00-289.99	15
290.00-299.99	10
300.00-319.99	8
320.00-379.99	3
	Total 70

En la tabla 2.9*a*) se conserva una misma amplitud de intervalo de clase, $10.00. Esto da por resultado que haya demasiadas clases vacías y que sea demasiado detallada en la parte superior de la escala de los salarios.

En la tabla 2.9*b*) se han evitado las clases vacías y el excesivo detalle empleando el intervalo abierto "$320 y más". La desventaja es que esta tabla no es útil para realizar ciertos cálculos matemáticos. Por ejemplo, no se puede determinar cuál es la cantidad total pagada como salarios semanalmente, ya que en "más de $320.00" puede haber individuos que ganen hasta $1 400.00 por semana.

En la tabla 2.9*c*) se emplea $20.00 como amplitud del intervalo de clase. La desventaja es que en el extremo inferior de la escala de salarios se pierde mucha información, en tanto que en el extremo superior de la escala, la tabla sigue siendo demasiado detallada.

En la tabla 2.9*d*) se emplean amplitudes desiguales de intervalos de clase. La desventaja es que se complican ciertos cálculos que puede desearse hacer después, lo que no ocurre cuando los intervalos de clase son de la misma amplitud. También, cuanto mayor sea la amplitud del intervalo de clase, mayor será el error de agrupamiento.

2.13 En la figura 2-8 se muestra un histograma, obtenido con EXCEL, con las distancias de la tabla 2.8. Las clases son 0 a 3, 3 a 6, 6 a 9, 9 a 12, 12 a 15 y 15 a 18. Los números que caigan en el límite superior de clase se cuentan dentro de esa clase, pero si caen en el límite inferior se cuentan dentro de la clase anterior.

a) ¿Cuáles son los valores (de la tabla 2.8) que pertenecen a la primera clase?

b) ¿Cuáles son los valores que pertenecen a la segunda clase?

c) ¿Cuáles son los valores que pertenecen a la tercera clase?

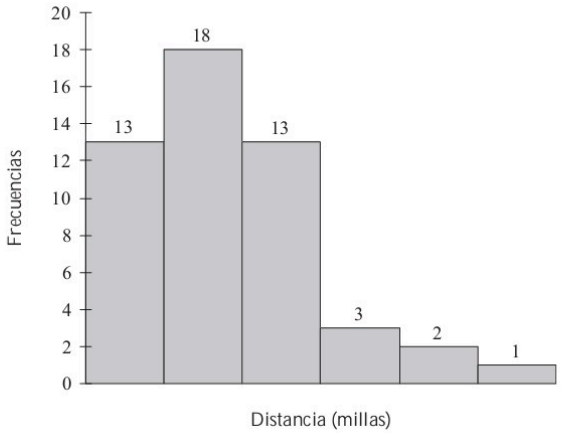


Figura 2-8 EXCEL, histograma con las distancias al Metropolitan College.

d) ¿Cuáles son los valores que pertenecen a la cuarta clase?
e) ¿Cuáles son los valores que pertenecen a la quinta clase?
f) ¿Cuáles son los valores que pertenecen a la sexta clase?

SOLUCIÓN

a) 0.9, 0.9, 1.0, 1.1, 1.4, 1.6, 1.9, 2.0, 2.0, 2.2, 2.4, 2.6, 3.0
b) 3.2, 3.3, 3.7, 3.8, 3.8, 3.9, 3.9, 4.0, 4.2, 4.3, 4.3, 4.4, 4.4, 4.6, 4.8, 4.8, 4.9, 5.0
c) 6.2, 6.5, 6.6, 7.0, 7.2, 7.7, 7.8, 8.0, 8.0, 8.0, 8.4, 8.7, 8.8
d) 10.0, 10.3, 10.3
e) 12.3, 12.6
f) 15.7

DISTRIBUCIONES DE FRECUENCIAS ACUMULADAS Y OJIVAS

2.14 A partir de la distribución de frecuencias dada en la tabla 2.5 del problema 2.3, construir: *a*) una distribución de frecuencias acumuladas, *b*) una distribución acumulada porcentual, *c*) una ojiva y *d*) una ojiva porcentual.

Tabla 2.10

Salarios	Frecuencias acumuladas	Distribución acumulada porcentual
Menos de $250.00	0	0.0
Menos de $260.00	8	12.3
Menos de $270.00	18	27.7
Menos de $280.00	34	52.3
Menos de $290.00	48	73.8
Menos de $300.00	58	89.2
Menos de $310.00	63	96.9
Menos de $320.00	65	100.0

SOLUCIÓN

a) y *b*) En la tabla 2.10 se muestran la distribución de frecuencias acumuladas y la distribución de frecuencia porcentual (o distribución de frecuencias acumuladas relativas).

Obsérvese que las entradas de la columna 2 se obtienen sumando las entradas sucesivas de la columna 2 de la tabla 2.5, así, $18 = 8 + 10$, $34 = 8 + 10 + 16$, etcétera.

Las entradas de la columna 3 se obtienen dividiendo cada una de las entradas de la columna anterior entre 65, la suma de todas las frecuencias, y expresando el resultado como porcentaje. Así, $34/65 = 0.523$, o 52.3%. Las entradas en esta columna también pueden obtenerse añadiendo entradas sucesivas de la columna 2 de la tabla 2.8. Así, $27.7 = 12.3 + 15.4$, $52.3 = 12.3 + 15.4 + 24.6$, etcétera.

c) y *d*) En la figura 2-9*a*) se muestra la ojiva (gráfica de frecuencias acumuladas porcentuales), y en la figura 2-9*b*) se presenta la ojiva porcentual (gráfica de frecuencias acumuladas relativas). Ambas son gráficas generadas con Minitab.

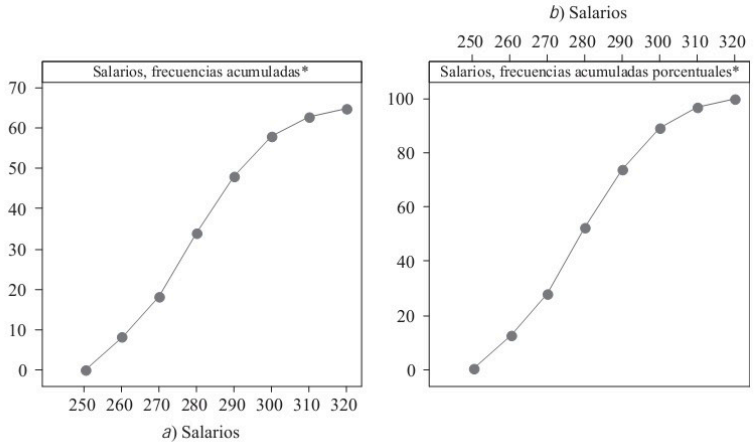


Figura 2-9 MINITAB, *a*) gráfica de frecuencias acumuladas y *b*) gráfica de frecuencias acumuladas porcentuales.

2.15 A partir de la distribución de frecuencias dada en la tabla 2.5 del problema 2.3, construir: *a*) una distribución de frecuencias "o más" y *b*) una ojiva "o más".

SOLUCIÓN

a) En la tabla 2.11, obsérvese que cada entrada de la columna 2 se obtiene sumando las entradas sucesivas de la columna 2 de la tabla 2.5, *empezando en la parte inferior* de la tabla 2.5; así, $7 = 2 + 5$, $17 = 2 + 5 + 10$, etc. Estas entradas también pueden obtenerse restando las entradas en la columna 2 de la tabla 2.10 del total de las frecuencias, 65; así, $57 = 65 - 8$, $47 = 65 - 18$, etcétera.

b) En la figura 2.10 se muestra la ojiva "o más".

Tabla 2.11

Salarios	Frecuencias acumuladas "o más"
$250.00 o más	65
$260.00 o más	57
$270.00 o más	47
$280.00 o más	31
$290.00 o más	17
$300.00 o más	7
$310.00 o más	2
$320.00 o más	0

2.16 A partir de las ojivas de las figuras 2-9 y 2-10 (problemas 2.14 y 2.15, respectivamente), estimar la cantidad de empleados que ganan: *a*) menos de $288.00 por semana, *b*) $296.00 o más por semana, *c*) por lo menos $263.00 por semana, pero menos de $275.00 por semana.

SOLUCIÓN

a) En la ojiva "menos de" de la figura 2-9 se traza una recta vertical que cruce la recta de los salarios en $288.00. Este punto cruza la ojiva en un punto cuyas coordenadas son (288, 45); por lo tanto, la cantidad de empleados que gana menos de $288.00 por semana es 45.

b) En la ojiva "o más" de la figura 2-10 se traza una recta vertical en $296.00. Esta recta cruza la ojiva en el punto (296, 11); por lo tanto, la cantidad de empleados que gana $296.00 o más por semana es 11 empleados.

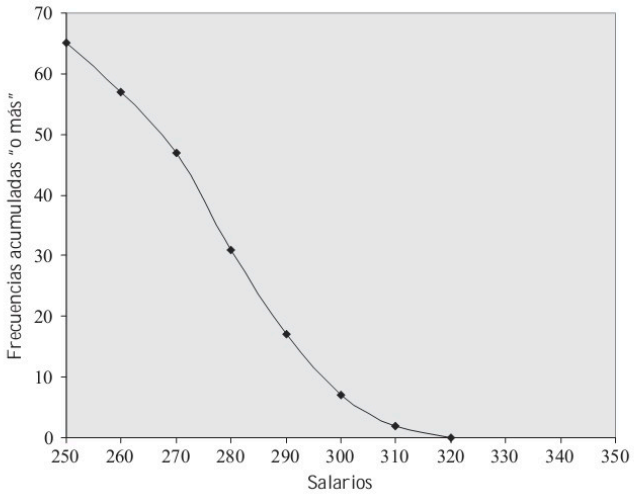


Figura 2-10 EXCEL, gráfica de frecuencias acumuladas "o más"

Esto también puede obtenerse a partir de la ojiva "menos de" de la figura 2-9. Trazando una recta en \$296.00, se encuentra que 54 empleados ganan menos de \$296.00 por semana; por lo tanto, $65 - 54 = 11$ empleados ganan \$296.00 o más por semana.

c) Utilizando la ojiva "menos de" de la figura 2-9, se tiene: cantidad de empleados buscada = cantidad de empleados que gana menos de \$275.00 por semana – cantidad de empleados que gana menos de \$263.00 por semana $= 26 - 11 = 15$.

Obsérvese que los resultados anteriores también pueden obtenerse mediante *interpolación* en la tabla de frecuencias acumuladas. Para el inciso *a*), por ejemplo, ya que \$288.00 está a 8/10 o 4/5 entre \$280.00 y \$290.00, la cantidad de empleados buscada debe estar a 4/5 entre 34 y 48 (ver tabla 2.10). Y 4/5 entre 34 y 48 es $4/5(48 - 34) = 11$. Por lo tanto, el número de empleados buscado es $3 + 11 = 45$.

2.17 Se lanzan cinco monedas 1 000 veces y en cada lanzamiento se anota el número de caras que se obtiene. En la tabla 2.12 se muestran la cantidades 0, 1, 2, 3, 4 y 5 de caras que se obtuvieron.

a) Graficar los datos de la tabla 2.12.
b) Elaborar una tabla en la que se dé el porcentaje de los lanzamientos en los que se obtuvo menos de 0, 1, 2, 3, 4, 5 y 6 caras.
c) Graficar los datos de la tabla del inciso *b*).

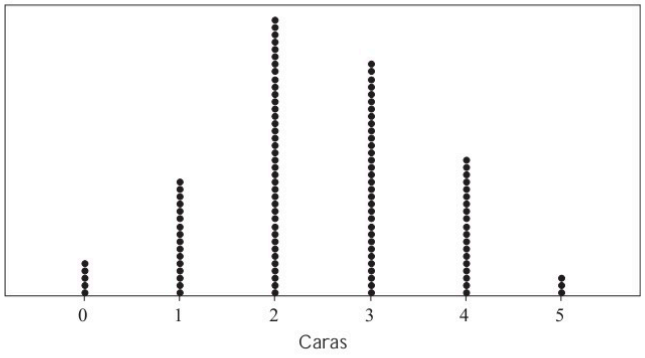
Tabla 2.12

Cantidad de caras	Cantidad de lanzamientos (frecuencias)
0	38
1	144
2	342
3	287
4	164
5	25
	Total 1 000

SOLUCIÓN

a) Estos datos se pueden mostrar gráficamente, ya sea como en la figura 2-11 o como en la figura 2-12.

Al parecer es más natural usar la figura 2-11, ya que la cantidad de caras no puede ser, por ejemplo, 1.5 o 3.2. A esta gráfica se le llama *gráfica de puntos* y se usa cuando los datos son discretos.



Cada símbolo (punto) representa hasta 9 observaciones.

Figura 2-11 MINITAB, gráfica de puntos con la cantidad de caras.

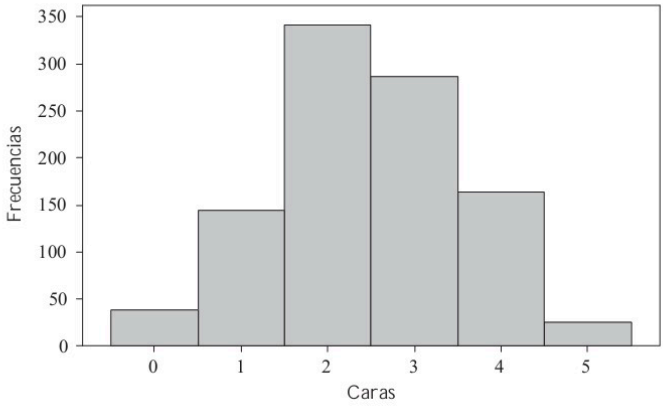


Figura 2-12 MINITAB, histograma de la cantidad de caras.

En la figura 2-12 se presenta un histograma de los datos. Obsérvese que toda el área del histograma corresponde a la frecuencia total, 1 000, como debe ser. Cuando se usa un histograma o el correspondiente polígono de frecuencias, se está tratando a los datos *como si* fueran continuos. Esto, como se verá más tarde, resulta útil. Obsérvese que ya en el problema 2.10 se usó un histograma y un polígono de frecuencias para datos discretos.

b) La tabla 2.13 es la requerida. Obsérvese que en esta tabla simplemente se da una distribución de las frecuencias acumuladas y una distribución de las frecuencias acumuladas porcentuales de la cantidad de caras. Hay que notar que las entradas "Menor de 1", "Menor de 2", etc., también podrían haber sido "Menor o igual a 0", "Menor o igual a 1", etcétera.

c) La gráfica pedida se puede representar como en la figura 2-13 o la figura 2-14.

La figura 2-13 es más natural para representar datos discretos, ya que el porcentaje de lanzamientos en el que se obtienen dos caras es igual al porcentaje en el que habrá menos de 1.75, 1.56 o 1.23 caras, es decir, es un mismo porcentaje (18.2%) el que corresponde a todos estos valores (lo que se indica por la línea horizontal).

Tabla 2.13

Número de caras	Número de lanzamientos (frecuencias acumuladas)	Cantidades porcentuales de lanzamientos (frecuencias acumuladas porcentuales)
Menos de 0	0	0.0
Menos de 1	38	3.8
Menos de 2	182	18.2
Menos de 3	524	52.4
Menos de 4	811	81.1
Menos de 5	975	97.5
Menos de 6	1 000	100.0

En la figura 2-14 se presenta la gráfica de frecuencias acumuladas, u ojiva; los datos se tratan como si fueran continuos.

Obsérvese que las figuras 2-13 y 2-14 corresponden, respectivamente, a las figuras 2-11 y 2-12 del inciso *a*).

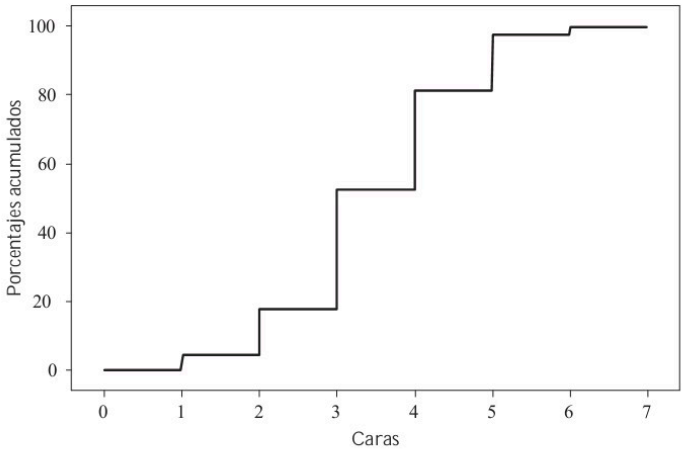


Figura 2-13 MINITAB, función escalonada.

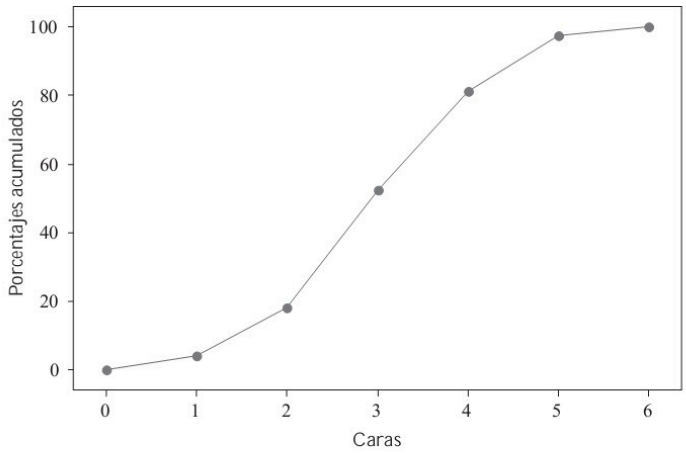


Figura 2-14 MINITAB, gráfica de frecuencias acumuladas.

CURVAS DE FRECUENCIAS Y OJIVAS SUAVIZADAS

2.18 Las muestras de poblaciones tienen histogramas y polígonos de frecuencias con ciertas formas. Si las muestras son muy grandes, los histogramas y los polígonos de frecuencias se aproximan a la distribución de la población. Considérense dos distribuciones de frecuencias poblacionales. *a*) Considérese una máquina que llena uniformemente envases de refresco con una cantidad entre 15.9 y 16.1 onzas. Trazar la curva de frecuencias y determinar qué porcentaje de los envases tiene más de 15.95 onzas. *b*) Considérense estaturas de mujeres. Estas estaturas tienen una distribución de frecuencias poblacional que es simétrica o en forma de campana, en la que el promedio es igual a 65 in y la desviación estándar es igual a 3 in. (La desviación estándar se estudia en un capítulo posterior.) ¿Qué porcentaje de las estaturas se encuentran entre 62 y 68 in, es decir, están a no más de una desviación estándar de la media? ¿Qué porcentaje se encuentra a no más de dos desviaciones estándar de la media? ¿Qué porcentaje se encuentra a no más de tres desviaciones estándar de la media?

SOLUCIÓN

En la figura 2-15 se muestra una curva de frecuencias uniforme. La región sombreada corresponde a los envases con más de 15.95 onzas. Obsérvese que la región abarcada por la curva de frecuencias tiene forma de rectángulo. El área bajo la curva de frecuencias está dada por largo $\times$ ancho, es decir $(16.10 - 15.90) \times 5 = 1$. El área de la región sombreada es $(16.10 - 15.95) \times 5 = 0.75$. Esto se interpreta como que 75% de los envases llenados tiene más de 15.95 onzas.

En la figura 2-16 se muestra una curva de frecuencias en forma de campana o simétrica. En esta figura se muestran en una región sombreada las estaturas a no más de una desviación estándar. Para calcular esta área es necesario hacer uso

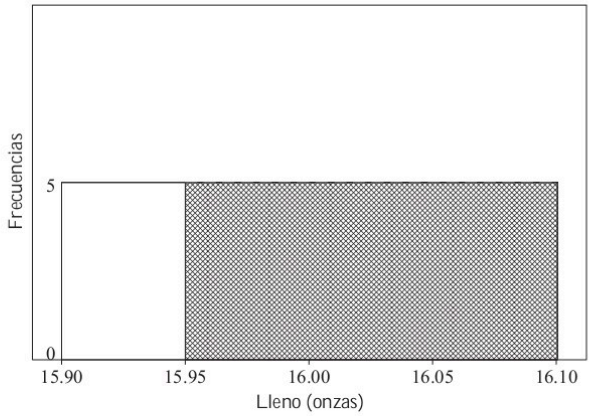


Figura 2-15 MINITAB, curva de frecuencias uniforme que muestra llenado a más de 15.95 onzas.

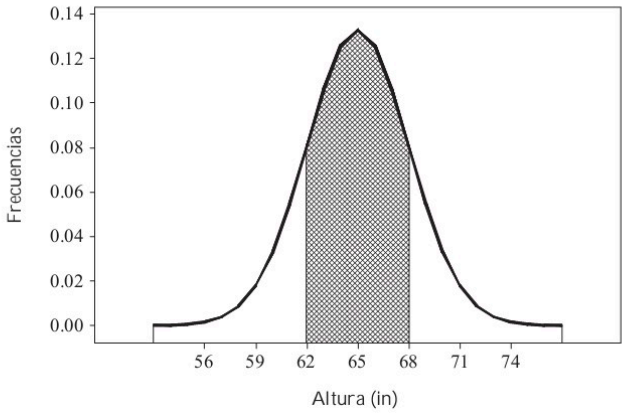


Figura 2-16 MINITAB, curva de frecuencias en forma de campana que muestra la altura entre 62 y 68 in y sus frecuencias.

del cálculo. El área comprendida a no más de una desviación estándar es aproximadamente 68% de toda el área bajo la curva. El área a no más de dos desviaciones estándar es aproximadamente 95% de toda el área bajo la curva. El área a no más de tres desviaciones estándar es aproximadamente 99.7% de toda el área bajo la curva.

En capítulos posteriores se verá más acerca de cómo encontrar áreas bajo estas curvas.

PROBLEMAS SUPLEMENTARIOS

2.19 *a*) Disponga los números 12, 56, 42, 21, 5, 18, 10, 3, 61, 34, 65 y 24 en una ordenación, y *b*) determine el rango.

2.20 En la tabla 2.14 se presenta una distribución de frecuencias de la cantidad de minutos por semana que ven televisión 400 estudiantes. De acuerdo con esta tabla, determinar:

a) El límite superior de la quinta clase.
b) El límite inferior de la octava clase.
c) La marca de clase de la séptima clase.
d) Las fronteras de clase de la última clase.
e) El tamaño del intervalo de clase.
f) La frecuencia de la cuarta clase.
g) La frecuencia relativa de la sexta clase.
h) El porcentaje de estudiantes que no ven televisión más de 600 minutos por semana.
i) El porcentaje de estudiantes que ven televisión 900 o más minutos por semana.
j) El porcentaje de estudiantes que ven televisión por lo menos 500 minutos por semana, pero menos de 1 000 minutos por semana.

Tabla 2.14

Tiempo (minutos)	Número de estudiantes
300-399	14
400-499	46
500-599	58
600-699	76
700-799	68
800-899	62
900-999	48
1 000-1 099	22
1 100-1 199	6

2.21 Elaborar: *a*) un histograma y *b*) un polígono de frecuencias para la distribución de frecuencias de la tabla 2.14.

2.22 Con los datos de la tabla 2.14 del problema 2.20, construir: *a*) una distribución de frecuencias relativas, *b*) un histograma de frecuencias relativas y *c*) un polígono de frecuencias relativas.

2.23 Con los datos de la tabla 2.14, construir: *a*) una distribución de frecuencias acumuladas, *b*) una distribución acumulada porcentual, *c*) una ojiva y *d*) una ojiva porcentual. (Obsérvese que a menos que se especifique otra cosa, una distribución acumulada es del tipo "menos que".)

2.24 Repetir el problema 2.23, pero para el caso en que las frecuencias acumuladas sean del tipo "o mayor".

2.25 Con los datos de la tabla 2.14, estimar el porcentaje de estudiantes que ven la televisión: *a*) menos de 560 minutos por semana, *b*) 970 o más minutos por semana y *c*) entre 620 y 890 minutos por semana.

2.26 El diámetro interno de las lavadoras producidas por una empresa se mide con una exactitud de milésimas de pulgada. Si las marcas de clase de la distribución de estos diámetros dados en pulgadas son 0.321, 0.324, 0.327, 0.330, 0.333 y 0.336, encontrar: *a*) la amplitud del intervalo de clase, *b*) las fronteras de clase y *c*) los límites de clase.

2.27 En la tabla siguiente se dan los diámetros en centímetros de una muestra de 60 balines fabricados en una empresa. Elaborar una distribución de frecuencias de los diámetros empleando los intervalos de clase adecuados.

1.738	1.729	1.743	1.740	1.736	1.741	1.735	1.731	1.726	1.737
1.728	1.737	1.736	1.735	1.724	1.733	1.742	1.736	1.739	1.735
1.745	1.736	1.742	1.740	1.728	1.738	1.725	1.733	1.734	1.732
1.733	1.730	1.732	1.730	1.739	1.734	1.738	1.739	1.727	1.735
1.735	1.732	1.735	1.727	1.734	1.732	1.736	1.741	1.736	1.744
1.732	1.737	1.731	1.746	1.735	1.735	1.729	1.734	1.730	1.740

2.28 Con los datos del problema 2.27, construir: *a*) un histograma, *b*) un polígono de frecuencias, *c*) una distribución de frecuencias relativas, *d*) un histograma de frecuencias relativas, *e*) un polígono de frecuencias relativas, *f*) una distribución de frecuencias acumuladas, *g*) una distribución acumulada porcentual, *h*) una ojiva, *i*) una ojiva porcentual.

2.29 Empleando los resultados del problema 2.28, determinar el porcentaje de balines cuyo diámetro: *a*) es mayor que 1.732 cm, *b*) no es mayor que 1.736 cm y *c*) está entre 1.730 y 1.738 cm. Comparar los resultados con los obtenidos directamente a partir de los datos en bruto del problema 2.27.

2.30 Repetir el problema 2.28 con los datos del problema 2.20.

2.31 De acuerdo con la Oficina de los Censos de Estados Unidos, en 1996 la población de este país era de 265 284 000. La tabla 2.15 da la distribución porcentual en los diversos grupos de edad.

a) ¿Cuál es la amplitud o el tamaño del segundo intervalo de clase? ¿Y la del cuarto intervalo de clase?
b) ¿Cuántos tamaños distintos de intervalos de clase hay?
c) ¿Cuántos intervalos de clase abiertos hay?
d) ¿Cómo se deberá escribir el último intervalo de clase de manera que su amplitud sea igual a la del penúltimo intervalo de clase?
e) ¿Cuál es la marca de clase del segundo intervalo de clase? ¿Y la del cuarto intervalo de clase?
f) ¿Cuáles son las fronteras de clase del cuarto intervalo de clase?
g) ¿Qué porcentaje de la población tiene 35 años o más? ¿Qué porcentaje de la población tiene 64 años o menos?
h) ¿Qué porcentaje de la población tiene entre 20 y 49 inclusive?
i) ¿Qué porcentaje de la población tiene más de 70 años?

2.32 *a*) ¿Por qué es imposible construir un histograma porcentual o un polígono de frecuencias con la distribución de la tabla 2.15?

b) ¿Cómo hay que modificar esta distribución para que se pueda construir un histograma porcentual o un polígono de frecuencias?
c) Usando la modificación del inciso *b*), construir estas gráficas.

Tabla 2.15

Grupo de edad en años	% de Estados Unidos
Menos de 5	7.3
5-9	7.3
10-14	7.2
15-19	7.0
20-24	6.6
25-29	7.2
30-34	8.1
35-39	8.5
40-44	7.8
45-49	6.9
50-54	5.3
55-59	4.3
60-64	3.8
65-74	7.0
75-84	4.3
85 o más	1.4

Fuente: U.S. Bureau of the Census, Current Population Reports.

2.33 Con relación a la tabla 2.15, supóngase que la población total es 265 millones y que la clase "menos de 5" comprende a niños menores de 1 año. Dar el número de individuos que hay en cada grupo, en millones, con una exactitud de una décima de millón.

2.34 *a*) Trazar un polígono de frecuencias porcentuales suavizado y una ojiva porcentual suavizada que correspondan a los datos de la tabla 2.14.

b) Empleando los resultados del inciso *a*), estimar la probabilidad de que un estudiante vea menos de 10 horas de televisión por semana.

c) Empleando los resultados del inciso *a*), estimar la probabilidad de que un estudiante vea 15 horas o más de televisión por semana.

d) Empleando los resultados del inciso *a*), estimar la probabilidad de que un estudiante vea menos de 5 horas de televisión por semana.

2.35 *a*) Lanzar 50 veces cuatro monedas y tabular la cantidad de caras que obtiene en cada lanzamiento.

b) Elaborar una distribución de frecuencias en la que se muestre la cantidad de lanzamientos en los que se obtuvo 0, 1, 2, 3 y 4 caras.

c) Elaborar la distribución porcentual correspondiente al inciso *b*).

d) Comparar los porcentajes obtenidos con los teóricos, 6.25%, 25%, 37.5%, 25% y 6.25% (proporcionales a 1, 4, 6, 4, y 1), que se obtienen por las reglas de la probabilidad.

e) Graficar las distribuciones de los incisos *b*) y *c*)

f) Trazar la ojiva porcentual correspondiente a los datos.

2.36 Repetir el problema 2.35 con 50 lanzamientos más de las cuatro monedas y ver si hay mayor coincidencia con lo que se espera teóricamente. Si no es así, dar los razonamientos que puedan explicar esas diferencias.

MEDIA, MEDIANA, MODA, Y OTRAS MEDIDAS DE TENDENCIA CENTRAL

3

ÍNDICES O SUBÍNDICES

El símbolo, X_j (que se lee "X subíndice j") representa cualquiera de los N *valores* $X_1, X_2, X_3, \ldots, X_N$ que puede tomar la variable X. A la letra j que aparece en X_j representando a cualquiera de los números 1, 2, 3, ..., N se le llama *subíndice* o *índice*. En lugar de j se puede usar, por supuesto, cualquier otra letra, *i, k, p, q* o *s*.

SUMATORIA

El símbolo $\sum_{j=1}^{N} X_j$ se emplea para denotar la suma de todas las X_j desde $j = 1$ hasta $j = N$; por definición,

$$\sum_{j=1}^{N} X_j = X_1 + X_2 + X_3 + \cdots + X_N$$

Cuando no puede haber confusión, esta suma se denota simplemente como $\sum X$, $\sum X_j$ o $\sum_j X_j$. El símbolo $\sum$ es la letra griega mayúscula *sigma* y denota suma.

EJEMPLO 1 $\displaystyle\sum_{j=1}^{N} X_j Y_j = X_1 Y_1 + X_2 Y_2 + X_3 Y_3 + \cdots + X_N Y_N$

EJEMPLO 2 $\displaystyle\sum_{j=1}^{N} aX_j = aX_1 + aX_2 + \cdots + aX_N = a(X_1 + X_2 + \cdots + X_N) = a\sum_{j=1}^{N} X_j$

donde a es una constante. O bien simplemente $\sum aX = a\sum X$.

EJEMPLO 3 Si a, b y c son cualesquiera constantes, entonces $\sum (aX + bY - cZ) = a\sum X + b\sum Y - c\sum Z$. Ver problema 3.3.

PROMEDIOS O MEDIDAS DE TENDENCIA CENTRAL

Un *promedio* es un valor típico o representativo de un conjunto de datos. Como estos valores típicos tienden a encontrarse en el centro de los conjuntos de datos, ordenados de acuerdo con su magnitud, a los promedios se les conoce también como *medidas de tendencia central.*

Se pueden definir varios tipos de promedios; los más usados son la *media aritmética,* la *mediana*, la *moda*, la *media geométrica* y la *media armónica.* Cada una de ellas tiene ventajas y desventajas de acuerdo con el tipo de datos y el propósito de su uso.

LA MEDIA ARITMÉTICA

La *media aritmética,* o brevemente la *media,* de un conjunto de N números $X_1, X_2, X_3, \ldots, X_N$ se denota así: $\bar{X}$ (que se lee "X barra") y está definida como

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\sum_{j=1}^{N} X_j}{N} = \frac{\sum X}{N} \tag{1}$$

EJEMPLO 4 La media aritmética de los números 8, 3, 5, 12 y 10 es

$$\bar{X} = \frac{8 + 3 + 5 + 12 + 10}{5} = \frac{38}{5} = 7.6$$

Si los números $X_1, X_2, \ldots, X_K$ se presentan $f_1, f_2, \ldots, f_K$ veces, respectivamente (es decir, se presentan con frecuencias $f_1, f_2, \ldots, f_K$), su media aritmética es

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + \cdots + f_KX_K}{f_1 + f_2 + \cdots + f_K} = \frac{\sum_{j=1}^{K} f_jX_j}{\sum_{j=1}^{K} f_j} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{N} \tag{2}$$

donde $N = \sum f$ es la *suma de las frecuencias* (es decir, la cantidad total de casos).

EJEMPLO 5 Si 5, 8, 6 y 2 se presentan con frecuencias 3, 2, 4 y 1, respectivamente, su media aritmética es

$$\bar{X} = \frac{(3)(5) + (2)(8) + (4)(6) + (1)(2)}{3 + 2 + 4 + 1} = \frac{15 + 16 + 24 + 2}{10} = 5.7$$

MEDIA ARITMÉTICA PONDERADA

Algunas veces, a los números $X_1, X_2, \ldots, X_K$ se les asignan ciertos *factores de ponderación* (o *pesos*) $w_1, w_2, \ldots, w_K$, que dependen del significado o importancia que se les asigne a estos números. En este caso, a

$$\bar{X} = \frac{w_1X_1 + w_2X_2 + \cdots + w_KX_k}{w_1 + w_2 + \cdots + w_K} = \frac{\sum wX}{\sum w} \tag{3}$$

se le llama *media aritmética ponderada.* Obsérvese la semejanza con la ecuación (*2*), la cual se puede considerar como una media aritmética ponderada con pesos $f_1, f_2, \ldots, f_K$.

EJEMPLO 6 Si en una clase, al examen final se le da el triple de valor que a los exámenes parciales y un estudiante obtiene 85 en el examen final, y 70 y 90 en los dos exámenes parciales, su puntuación media es

$$\bar{X} = \frac{(1)(70) + (1)(90) + (3)(85)}{1 + 1 + 3} = \frac{415}{5} = 83$$

PROPIEDADES DE LA MEDIA ARITMÉTICA

1. En un conjunto de números, la suma algebraica de las desviaciones de estos números respecto a su media aritmética es cero.

EJEMPLO 7 Las desviaciones de los números 8, 3, 5, 12 y 10 de su media aritmética, 7.6, son $8 - 7.6$, $3 - 7.6$, $5 - 7.6$, $12 - 7.6$ y $10 - 7.6$ o bien 0.4, −4.6, −2.6, 4.4 y 2.4, cuya suma algebraica es $0.4 - 4.6 - 2.6 + 4.4 + 2.4 = 0$.

2. En un conjunto de números X_j, la suma de los cuadrados de sus desviaciones respecto a un número a es un mínimo si y sólo si $a = \bar{X}$ (ver el problema 4.27).
3. Si la media de f_1 números es m_1, la media de f_2 números es $m_2, \ldots$, la media de f_k números es m_k, entonces la media de todos estos números es

$$\bar{X} = \frac{f_1 m_1 + f_2 m_2 + \cdots + f_K m_K}{f_1 + f_2 + \cdots + f_K} \tag{4}$$

   es decir, una media aritmética ponderada de todas las medias (véase el problemas 3.12).
4. Si se *cree* o *se supone* que un número A (que puede ser cualquier número) es la *media aritmética* y si $d_j = X_j - A$ son las desviaciones de X_j de A, entonces las ecuaciones (*1*) y (*2*) se convierten, respectivamente, en

$$\bar{X} = A + \frac{\sum_{j=1}^{N} d_j}{N} = A + \frac{\sum d}{N} \tag{5}$$

$$\bar{X} = A + \frac{\sum_{j=1}^{K} f_j d_j}{\sum_{j=1}^{K} f_j} = A + \frac{\sum fd}{N} \tag{6}$$

   donde $N = \sum_{j=1}^{N} f_j = \sum f$. Obsérvese que las fórmulas (*5*) y (*6*) se resumen en la ecuación $\bar{X} = A + \bar{d}$ (ver problema 3.18).

CÁLCULO DE LA MEDIA ARITMÉTICA PARA DATOS AGRUPADOS

Cuando se presentan los datos en una distribución de frecuencias, se considera que todos los datos que caen en un intervalo de clase dado coinciden con la marca o punto medio del intervalo. Para datos agrupados, interpretando a las X_j como las marcas de clase, a las f_j como las correspondientes frecuencias de clase, a A como cualquier marca de clase supuesta y $d_j = X_j - A$ como la desviación de X_j respecto de A, las fórmulas (*2*) y (*6*) son válidas.

A los cálculos empleando las fórmulas (*2*) y (*6*) se les suele conocer como *método largo* y *método abreviado*, respectivamente (ver los problemas 3.15 y 3.20).

Si todos los intervalos de clase son de una misma amplitud c, las desviaciones $d_j = X_j - A$ se pueden expresar como cu_j, donde u_j puede tener valores enteros positivos o negativos o cero (es decir, 0, ±1, ±2, ±3, ...) con lo que la fórmula (*6*) se convierte en

$$\bar{X} = A + \left(\frac{\sum_{j=1}^{K} f_j u_j}{N} \right) = A + \left(\frac{\sum fu}{N} \right) c \tag{7}$$

lo que es equivalente a la ecuación $\bar{X} = A + c\bar{u}$ (ver problema 3.21). A esta ecuación se le conoce como *método codificado* para calcular la media. Es un método muy breve recomendado para datos agrupados cuando los intervalos de clase tienen todos la misma amplitud (ver problemas 3.22 y 3.23). Obsérvese que en el método codificado los valores de la variable X *se transforman* en valores de la variable u de acuerdo con $X = A + cu$.

LA MEDIANA

La *mediana* de un conjunto de números acomodados en orden de magnitud (es decir, en una ordenación) es el valor central o la media de los dos valores centrales.

EJEMPLO 8 La mediana del conjunto de números 3, 4, 5, 6, 8, 8, 8 y 10 es 6.

EJEMPLO 9 La mediana del conjunto de números 5, 5, 7, 9, 11, 12, 15 y 18 es $\frac{1}{2}(9 + 11) = 10$.

En datos agrupados, la mediana se obtiene por interpolación, como se expresa por la fórmula

$$\text{Mediana} = L_1 + \left(\frac{\frac{N}{2} - (\sum f)_1}{f_{\text{mediana}}} \right) c \tag{8}$$

donde $L_1 =$ frontera inferior de la clase mediana (es decir, de la clase que contiene la mediana)
$N =$ número de datos (es decir, la frecuencia total)
$(\sum f)_1 =$ suma de las frecuencias de todas las clases anteriores a la clase mediana
$f_{\text{mediana}} =$ frecuencia de la clase mediana
$c =$ amplitud del intervalo de la clase mediana

Geométricamente, la mediana es el valor de X (abscisa) que corresponde a una recta vertical que divide al histograma en dos partes que tienen la misma área. A este valor de X se le suele denotar $\tilde{X}$.

LA MODA

La *moda* de un conjunto de números es el valor que se presenta con más frecuencia; es decir, es el valor más frecuente. Puede no haber moda y cuando la hay, puede no ser única.

EJEMPLO 10 La moda del conjunto 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12 y 18 es 9.

EJEMPLO 11 El conjunto 3, 5, 8, 10, 12, 15 y 16 no tiene moda.

EJEMPLO 12 El conjunto 2, 3, 4, 4, 4, 5, 5, 7, 7, 7 y 9 tiene dos modas, 4 y 7, por lo que se le llama *bimodal*.

A una distribución que sólo tiene una moda se le llama *unimodal*.

En el caso de datos agrupados, para los que se ha construido una curva de frecuencia que se ajuste a los datos, la moda es el valor (o los valores) de X que corresponden al punto (o puntos) máximos de la curva. A este valor de X se le suele denotar $\hat{X}$.

En una distribución de frecuencia o en un histograma la moda se puede obtener mediante la fórmula siguiente:

$$\text{Moda} = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) c \tag{9}$$

donde $L_1 =$ frontera inferior de la clase modal (es decir, de la clase que contiene la moda)
$\Delta_1 =$ exceso de frecuencia modal sobre la frecuencia en la clase inferior inmediata
$\Delta_2 =$ exceso de frecuencia modal sobre la frecuencia en la clase superior inmediata
$c =$ amplitud del intervalo de la clase modal

RELACIÓN EMPÍRICA ENTRE LA MEDIA, LA MEDIANA Y LA MODA

En las curvas de frecuencias unimodales que son ligeramente sesgadas (asimétricas), se tiene la relación empírica siguiente:

$$\text{Media} - \text{moda} = 3(\text{media} - \text{mediana}) \tag{10}$$

En las figuras 3-1 y 3-2 se muestran las posiciones relativas de la media, la mediana y la moda en curvas de frecuencias sesgadas a la derecha o a la izquierda, respectivamente. En las curvas simétricas, la media, la mediana y la moda coinciden.

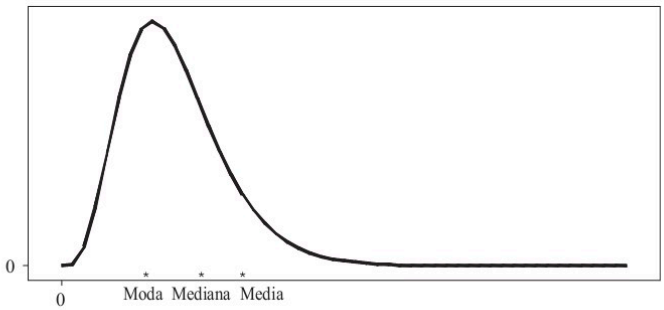


Figura 3-1 Posiciones relativas de la media, la mediana y la moda en curvas de frecuencias sesgadas a la derecha.

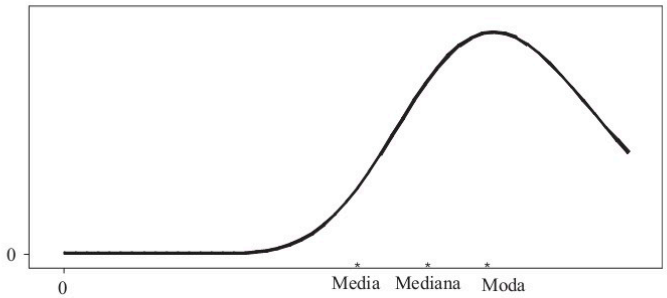


Figura. 3-2 Posiciones relativas de la media, la mediana y la moda en curvas de frecuencias sesgadas a la izquierda.

LA MEDIA GEOMÉTRICA *G*

La media geométrica *G* de *N* números positivos $X_1, X_2, X_3, \ldots, X_N$ es la raíz *n*-ésima del producto de los números:

$$G = \sqrt[N]{X_1 X_2 X_3 \cdots X_N} \tag{11}$$

EJEMPLO 13 La media geométrica de los números 2, 4 y 8 es $G = \sqrt[3]{(2)(4)(8)} = \sqrt[3]{64} = 4$.

G se puede calcular empleando logaritmos (ver problema 3.35) o usando una calculadora. Para la media geométrica de datos agrupados, ver los problemas 3.36 y 3.91.

LA MEDIA ARMÓNICA *H*

La media armónica *H* de un conjunto de *N* números $X_1, X_2, X_3, \ldots, X_N$ es el recíproco de la media aritmética de los recíprocos de los números:

$$H = \frac{1}{\dfrac{1}{N}\displaystyle\sum_{j=1}^{N}\frac{1}{X_j}} = \frac{N}{\sum \dfrac{1}{X}} \tag{12}$$

En la práctica es más fácil recordar que

$$\frac{1}{H} = \frac{\sum \frac{1}{X}}{N} = \frac{1}{N}\sum \frac{1}{X} \tag{13}$$

EJEMPLO 14 La media armónica de los números 2, 4 y 8 es

$$H = \frac{3}{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}} = \frac{3}{\frac{7}{8}} = 3.43$$

Para la media armónica de datos agrupados ver los problemas 3.99 y 3.100.

RELACIÓN ENTRE LAS MEDIAS ARITMÉTICA, GEOMÉTRICA Y ARMÓNICA

La media geométrica de un conjunto de números positivos $X_1, X_2, \ldots, X_N$ es menor o igual que su media aritmética, pero mayor o igual que su media armónica. En símbolos,

$$H \leq G \leq \bar{X} \tag{14}$$

La igualdad es válida sólo cuando todos los números $X_1, X_2, \ldots, X_N$ son idénticos.

EJEMPLO 15 La media aritmética de los números 2, 4 y 8 es 4.67, su media geométrica es 4 y su media armónica es 3.43.

LA RAÍZ CUADRADA MEDIA

La raíz cuadrada media (RCM) o *media cuadrática* de un conjunto de números $X_1, X_2, \ldots, X_N$ suele denotarse $\sqrt{\overline{X^2}}$ y se define

$$\text{RCM} = = \sqrt{\overline{X^2}} = \sqrt{\frac{\sum_{j=1}^{N} X_j^2}{N}} = \sqrt{\frac{\sum X^2}{N}} \tag{15}$$

Este tipo de promedio suele usarse en aplicaciones físicas.

EJEMPLO 16 La raíz cuadrada media del conjunto 1, 3, 4, 5, y 7 es

$$\sqrt{\frac{1^2+3^2+4^2+5^2+7^2}{5}} = \sqrt{20} = 4.47$$

CUARTILES, DECILES Y PERCENTILES

En un conjunto de datos en el que éstos se hallan ordenados de acuerdo con su magnitud, el valor de en medio (o la media aritmética de los dos valores de en medio), que divide al conjunto en dos partes iguales, es la mediana. Continuando con esta idea se puede pensar en aquellos valores que dividen al conjunto de datos en cuatro partes iguales. Estos valores, denotados Q_1, Q_2 y Q_3 son el primero, segundo y tercer *cuartiles*, respectivamente; el valor Q_2 coincide con la mediana.

De igual manera, los valores que dividen al conjunto en diez partes iguales son los *deciles* y se denotan D_1, $D_2, \ldots, D_9$, y los valores que dividen al conjunto en 100 partes iguales son los percentiles y se les denota P_1, $P_2, \ldots, P_{99}$. El quinto decil y el percentil 50 coinciden con la mediana. Los percentiles 25 y 75 coinciden con el primero y tercer cuartiles, respectivamente.

A los cuartiles, deciles, percentiles y otros valores obtenidos dividiendo al conjunto de datos en partes iguales se les llama en conjunto *cuantiles*. Para el cálculo de estos valores cuando se tienen datos agrupados ver los problemas 3.44 a 3.46.

EJEMPLO 17 Utilizar EXCEL para hallar Q_1, Q_2, Q_3, D_9 y P_{95}, en la muestra siguiente de puntuaciones.

88	45	53	86	33	86	85	30	89	53	41	96	56	38	62
71	51	86	68	29	28	47	33	37	25	36	33	94	73	46
42	34	79	72	88	99	82	62	57	42	28	55	67	62	60
96	61	57	75	93	34	75	53	32	28	73	51	69	91	35

Para encontrar el primer cuartil, ingrese los datos en los primeros 60 renglones de la columna A de la hoja de cálculo de EXCEL. Después, dé el comando =PERCENTILE(A1:A60,0.25). EXCEL da el valor 37.75. Se encuentra que 15 de los 60 valores, o el 25%, son menores que 37.75. De igual manera =PERCENTILE(A1:A60,0.5) da 57, =PERCENTILE(A1:A60,0.75) da 76, =PERCENTILE(A1:A60,0.9) da 89.2, =PERCENTILE(A1:A60,0.95) da 94.1. EXCEL da los cuartiles, deciles y percentiles expresados como percentiles.

A continuación se describe un algoritmo que suele emplearse para hallar cuartiles, deciles y percentiles. Primero se ordenan los datos del ejemplo 17 de acuerdo con su magnitud; el resultado es:

Puntuaciones de examen

25	28	28	28	29	30	32	33	33	33	34	34	35	36	37
38	41	42	42	45	46	47	51	51	53	53	53	55	56	57
57	60	61	62	62	62	67	68	69	71	72	73	73	75	75
79	82	85	86	86	86	88	88	89	91	93	94	96	96	99

Supóngase que se quiere encontrar el primer cuartil (que es el percentil 25). Se calcula $i = np/100 = 60(25)/100 = 15$. Como 15 es un número entero, se saca el promedio de los datos en las posiciones 15 y 16 de los datos ordenados de menor a mayor. Es decir, se promedian 37 y 38 y se obtiene 37.5 como primer cuartil ($Q_1 = 37.5$). Para hallar el percentil 93, se calcula $np/100 = 60(93)/100$ y se obtiene 55.8. Como este número no es un entero, se redondea hacia arriba y se obtiene 56. El número que ocupa la posición 56 en los datos ordenados es 93 y $P_{93} = 93$. El comando de EXCEL =PERCENTILE(A1:A60,0.93) da 92.74. Obsérvese que con EXCEL no se obtienen los mismos valores para los percentiles, pero sí valores cercanos. A medida que los conjuntos de datos son mayores, tienden a obtenerse los mismos valores.

SOFTWARE Y MEDIDAS DE TENDENCIA CENTRAL

Todos los paquetes de software utilizados en este libro dan las estadísticas descriptivas vistas en esta sección. A continuación se presentan los resultados que se obtienen con estos cinco paquetes empleando las puntuaciones de examen del ejemplo 17.

EXCEL

Seleccionando la secuencia "Tools ⇒ Data Analysis ⇒ Descriptive Statistics", se obtienen las medidas de tendencia central mediana, media y moda, así como varias medidas de dispersión.

Media	59.16667
Error típico	2.867425
Mediana	57
Moda	28
Desviación estándar	22.21098
Varianza de la muestra	493.3277
Curtosis	–1.24413
Coeficiente de asimetría	0.167175
Rango	74
Mínimo	25
Máximo	99
Suma	3 550
Cuenta	60

MINITAB

Si se selecciona la secuencia "Stat ⇒ Basic Statistics ⇒ Display Descriptive Statistics", como resultado se obtiene:

Estadística descriptiva: calificación de examen

Variable	N	N*	Media	SE media	Desv est	Mínimo	Q1	Mediana	Q3	Máxima
Punt examen	60	0	59.17	2.87	22.21	25.00	37.25	57.00	78.00	99.00

SPSS

Si se selecciona la secuencia "Analyze ⇒ Descriptive Statistics ⇒ Descriptives", como resultado se obtiene:

Estadística descriptiva

	N	Mínimo	Máximo	Media	Desviación estándar
Puntuación de examen	60	25.00	99.00	59.1667	22.21098
N válida	60				

SAS

Si se selecciona la secuencia **"Solutions ⇒ Análisis ⇒ Analyst"** y los datos se leen como un archivo, seleccionando la secuencia **"Statistics ⇒ Descriptive ⇒ Summary Statistics"**, se obtiene como resultado:

```
                    The MEANS Procedure
                    Analysis Variable : Testscr

        Mean          Std Dev       N          Minimum       Maximum
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
  59.1666667     22.2109811      60        25.0000000    99.0000000
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
```

STATISTIX

Si se selecciona la secuencia "Statistics ⇒ Summary Statistics ⇒ Descriptive Statistics" del paquete STATISTIX, como resultado se obtiene:

```
Statistix 8.0
Descriptive Staistics

                Testscore
N                      60
Mean               59.167
SD                 22.211
Minimum            25.000
1st Quarti         37.250
3rd Quarti         78.000
Maximum            99.000
```

PROBLEMAS RESUELTOS

SUMATORIA

3.1 Escribir los términos de cada una de las sumas siguientes:

a) $\sum_{j=1}^{6} X_j$ *c)* $\sum_{j=1}^{N} a$ *e)* $\sum_{j=1}^{3} (X_j - a)$

b) $\sum_{j=1}^{4} (Y_j - 3)^2$ *d)* $\sum_{k=1}^{5} f_k X_k$

SOLUCIÓN

a) $X_1 + X_2 + X_3 + X_4 + X_5 + X_6$

b) $(Y_1 - 3)^2 + (Y_2 - 3)^2 + (Y_3 - 3)^2 + (Y_4 - 3)^2$

c) $a + a + a + \cdots + a = Na$

d) $f_1X_1 + f_2X_2 + f_3X_3 + f_4X_4 + f_5X_5$

e) $(X_1 - a) + (X_2 - a) + (X_3 - a) = X_1 + X_2 + X_3 - 3a$

3.2 Expresar cada una de las sumas siguientes empleado el símbolo de sumatoria.

a) $X_1^2 + X_2^2 + X_3^2 + \cdots + X_{10}^2$

b) $(X_1 + Y_1) + (X_2 + Y_2) + \cdots + (X_8 + Y_8)$

c) $f_1X_1^3 + f_2X_2^3 + \cdots + f_{20}X_{20}^3$

d) $a_1b_1 + a_2b_2 + a_3b_3 + \cdots + a_Nb_N$

e) $f_1X_1Y_1 + f_2X_2Y_2 + f_3X_3Y_3 + f_4X_4Y_4$

SOLUCIÓN

a) $\sum_{j=1}^{10} X_j^2$ *c)* $\sum_{j=1}^{20} f_jX_j^3$ *e)* $\sum_{j=1}^{4} f_jX_jY_j$

b) $\sum_{j=1}^{8} (X_j + Y_j)$ *d)* $\sum_{j=1}^{N} a_jb_j$

3.3 Probar que $\sum_{j=1}^{N} (aX_j + bY_j - cZ_j) = a\sum_{j=1}^{N} X_j + b\sum_{j=1}^{N} Y_j - c\sum_{j=1}^{N} Z_j$, donde a, b y c son constantes.

SOLUCIÓN

$$\begin{aligned}\sum_{j=1}^{N} (aX_j + bY_j - cZ_j) &= (aX_1 + bY_1 - cZ_1) + (aX_2 + bY_2 - cZ_2) + \cdots + (aX_N + bY_N - cZ_N) \\ &= (aX_1 + aX_2 + \cdots + aX_N) + (bY_1 + bY_2 + \cdots + bY_N) - (cZ_1 + cZ_2 + \cdots + cZ_N) \\ &= a(X_1 + X_2 + \cdots + X_N) + b(Y_1 + Y_2 + \cdots + Y_N) - c(Z_1 + Z_2 + \cdots + Z_N) \\ &= a\sum_{j=1}^{N} X_j + b\sum_{j=1}^{N} Y_j - c\sum_{j=1}^{N} Z_j\end{aligned}$$

o brevemente, $\sum (aX + bY - cZ) = a\sum X + b\sum Y - c\sum Z$.

3.4 Dos variables, X y Y toman los valores $X_1 = 2$, $X_2 = -5$, $X_3 = 4$, $X_4 = -8$ y $Y_1 = -3$, $Y_2 = -8$, $Y_3 = 10$, $Y_4 = 6$, respectivamente. Calcular *a*) $\sum X$, *b*) $\sum Y$, *c*) $\sum XY$, *d*) $\sum X^2$, *e*) $\sum Y^2$, *f*) $(\sum X)(\sum Y)$, *g*) $\sum XY^2$ y *h*) $\sum (X + Y)(X - Y)$.

SOLUCIÓN

Obsérvese que en todos los casos se ha omitido en X y Y el subíndice j y que la $\sum$ se entiende como $\sum_{j=1}^{4}$. Por lo tanto, por ejemplo $\sum X$ es abreviación de $\sum_{j=1}^{4} X_j$.

a) $\sum X = (2) + (-5) + (4) + (-8) = 2 - 5 + 4 - 8 = -7$

b) $\sum Y = (-3) + (-8) + (10) + (6) = -3 - 8 + 10 + 6 = 5$

c) $\sum XY = (2)(-3) + (-5)(-8) + (4)(10) + (-8)(6) = -6 + 40 + 40 - 48 = 26$

d) $\sum X^2 = (2)^2 + (-5)^2 + (4)^2 + (-8)^2 = 4 + 25 + 16 + 64 = 109$

e) $\sum Y^2 = (-3)^2 + (-8)^2 + (10)^2 + (6)^2 = 9 + 64 + 100 + 36 = 209$

f) $(\sum X)(\sum Y) = (-7)(5) = -35$, de acuerdo con los incisos *a*) y *b*). Obsérvese que $(\sum X)(\sum Y) \neq \sum XY$.

g) $\sum XY^2 = (2)(-3)^2 + (-5)(-8)^2 + (4)(10)^2 + (-8)(6)^2 = -190$

h) $\sum (X + Y)(X - Y) = \sum (X^2 - Y^2) = \sum X^2 - \sum Y^2 = 109 - 209 = -100$, de acuerdo con los incisos *d*) y *e*).

3.5 En una nota de *USA Today* se informa que el promedio de impuestos, per cápita, recolectados en 2005, en todo Estados Unidos, fue de $2 189.84. Esta cantidad se desglosa así: ventas e ingresos, $1051.42; ingreso, $875.23; licencias, $144.33; otros, $80.49, y propiedades, $38.36. Usando EXCEL, demostrar que la suma es igual a $2 189.84.

SOLUCIÓN

Obsérvese que la expresión =sum(A1:A5) es equivalente a $\sum_{j=1}^{5} X_j$.

1 051.42	ventas e ingresos
875.23	ingreso
144.33	licencias
80.49	otros
38.36	propiedades
2 189.83	=sum(A1:A5)

LA MEDIA ARITMÉTICA

3.6 Las calificaciones de un estudiante en seis exámenes fueron 84, 91, 72, 68, 87 y 78. Hallar la media aritmética de estas calificaciones.

SOLUCIÓN

$$\bar{X} = \frac{\sum X}{N} = \frac{84 + 91 + 72 + 68 + 87 + 78}{6} = \frac{480}{6} = 80$$

El término *promedio* suele emplearse como sinónimo de *media aritmética*. Sin embargo, estrictamente hablando, esto no es correcto, ya que además de la media hay otros promedios.

3.7 Un científico mide diez veces el diámetro de un cilindro y obtiene los valores 3.88, 4.09, 3.92, 3.97, 4.02, 3.95, 4.03, 3.92, 3.98 y 4.06 centrímetros (cm). Hallar la media aritmética de estas mediciones.

SOLUCIÓN

$$\bar{X} = \frac{\sum X}{N} = \frac{3.88 + 4.09 + 3.92 + 3.97 + 4.02 + 3.95 + 4.03 + 3.92 + 3.98 + 4.06}{10} = \frac{39.82}{10} = 3.98\,\text{cm}$$

3.8 En el siguiente resultado obtenido con MINITAB se muestra la cantidad de tiempo por semana que 30 personas estuvieron empleando en Internet, así como la media de estas cantidades. ¿Podría decirse que este promedio es típico de las 30 cantidades?

```
MTB > print c1
```

Muestra de datos

```
tiempo
3   4   4   5   5   5   5   5   5   6
6   6   6   7   7   7   7   7   8   8
9  10  10  10  10  10  10  12  55  60
```

```
MTB > mean c1
```

Media de la columna

```
Mean of time = 10.400
```

SOLUCIÓN

Esta media de 10.4 horas no es típica de estas cantidades. Obsérvese que 21 de estas cantidades son de un solo dígito y que la media es 10.4 horas. Una gran desventaja de la media es que es fuertemente afectada por valores atípicos (o valores extremos.)

3.9 Encontrar la media aritmética de los números 5, 3, 6, 5, 4, 5, 2, 8, 6, 5, 4, 8, 3, 4, 5, 4, 8, 2, 5 y 4.

SOLUCIÓN

Primer método

$$\bar{X} = \frac{\sum X}{N} = \frac{5+3+6+5+4+5+2+8+6+5+4+8+3+4+5+4+8+2+5+4}{20} = \frac{96}{20} = 4.8$$

Segundo método

Hay las siguientes cantidades: seis 5, dos 3, dos 6, cinco 4, dos 2 y tres 8. Por lo tanto

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{N} = \frac{(6)(5) + (2)(3) + (2)(6) + (5)(4) + (2)(2) + (3)(8)}{6+2+2+5+2+3} = \frac{96}{20} = 4.8$$

3.10 De 100 números, 20 fueron 4, 40 fueron 5, 30 fueron 6 y los restantes fueron 7. Encuéntrese la media aritmética de estos números.

SOLUCIÓN

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{N} = \frac{(20)(4) + (40)(5) + (30)(6) + (10)(7)}{100} = \frac{530}{100} = 5.30$$

3.11 Las calificaciones finales de un estudiante en matemáticas, física, inglés e higiene son, respectivamente, 82, 86, 90 y 70. Si los créditos en cada uno de estos cursos son 3, 5, 3 y 1, determinar la correspondiente calificación promedio.

SOLUCIÓN

Se emplea la media aritmética ponderada, en donde los pesos que corresponden a cada puntuación son los créditos que les corresponden. Así,

$$\bar{X} = \frac{\sum wX}{\sum w} = \frac{(3)(82) + (5)(86) + (3)(90) + (1)(70)}{3 + 5 + 3 + 1} = 85$$

3.12 En una empresa en la que hay 80 empleados, 60 ganan \$10.00 por hora y 20 ganan \$13.00 por hora.

a) Determinar el sueldo medio por hora.
b) En el inciso *a*), ¿se obtiene la misma respuesta si los 60 empleados tienen un salario promedio de \$10.00 por hora? Probar la respuesta.
c) ¿Se considera que este salario medio por hora es representativo?

SOLUCIÓN

a)

$$\bar{X} = \frac{\sum fX}{N} = \frac{(60)(\$10.00) + (20)(\$13.00)}{60 + 20} = \$10.75$$

b) Sí, el resultado es el mismo. Para probar esto supóngase que la media de f_1 números es m_1 y que la media de f_2 números es m_2. Hay que demostrar que la media de todos estos números es

$$\bar{X} = \frac{f_1 m_1 + f_2 m_2}{f_1 + f_2}$$

Sea M_1 la suma de los f_1 números y M_2 la suma de los f_2 números. Entonces, por definición de media aritmética,

$$m_1 = \frac{M_1}{f_1} \qquad \text{y} \qquad m_2 = \frac{M_2}{f_2}$$

o $M_1 = f_1 m_1$ y $M_2 = f_2 m_2$. Como todos los $(f_1 + f_2)$ números suman $(M_1 + M_2)$, la media aritmética de todos estos números es

$$\bar{X} = \frac{M_1 + M_2}{f_1 + f_2} = \frac{f_1 m_1 + f_2 m_2}{f_1 + f_2}$$

como se deseaba. Este resultado se puede ampliar fácilmente.

c) Se puede decir que \$10.75 es un salario "representativo" por hora en el sentido de que la mayor parte de los empleados gana \$10.00 por hora, lo que no se aleja mucho de \$10.75 por hora. Se debe recordar que siempre que se resuman datos numéricos en un solo dato (como en un promedio) es posible que se cometa algún error. Sin embargo, el resultado desorienta tanto como en el problema 3.8

En realidad, para tener una mejor idea se debe dar una estimación de la "dispersión" o "variación" de los datos con respecto a la media. A esto se le llama *dispersión* de los datos. En el capítulo 4 se dan varias medidas de dispersión.

3.13 Los pesos medio de cuatro grupos de estudiantes que constan de 15, 20, 10 y 18 individuos son 162, 148, 153 y 140 libras, respectivamente. Encuentre el peso medio de todos los estudiantes.

SOLUCIÓN

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{(15)(162) + (20)(148) + (10)(153) + (18)(140)}{15 + 20 + 10 + 18} = 150\,\text{lb}$$

3.14 El ingreso medio anual de trabajadores agrícolas y no agrícolas es \$25 000 y \$35 000, respectivamente; ¿el ingreso medio anual de los dos grupos será \$30 000?

SOLUCIÓN

Sería \$30 000 únicamente si la cantidad de trabajadores agrícolas y no agrícolas fuese la misma. Para determinar el verdadero ingreso medio anual se necesita saber cuál es la cantidad relativa de trabajadores en cada grupo. Supóngase que 10% de los trabajadores son trabajadores agrícolas. En ese caso la media será $(0.10)(25\,000) + (0.90)(35\,000) = \$34\,000$. Si la cantidad de trabajadores de ambos tipos es la misma, la media será $(0.50)(25\,000) + (0.50)(35\,000) = \$30\,000$.

3.15 Usando la distribución de frecuencias de las estaturas que se presenta en la tabla 2.1, hallar la estatura media de los 100 estudiantes de la universidad XYZ.

SOLUCIÓN

En la tabla 3.1 se presentan los datos organizados para hacer los cálculos. Obsérvese que como estatura de los estudiantes que miden de 60 a 62 pulgadas (in), de 63 a 65 in, etc., se toman 61 in, 64 in, etc., respectivamente. Entonces, el problema se reduce a encontrar la estatura media de 100 estudiantes si 5 tienen una estatura de 61 in, 18 tienen una estatura de 64 in, etcétera.

Estos cálculos pueden resultar tediosos, en especial en los casos en que los números son grandes y se tienen muchas clases. Existen técnicas abreviadas para reducir el trabajo; ver los problemas 3.20 y 3.22.

Tabla 3.1

Estatura (in)	Marcas de clase (X)	Frecuencias (f)	fX
60-62	61	5	305
63-65	64	18	1 152
66-68	67	42	2 814
69-71	70	27	1 890
72-74	73	8	584
		$N = \sum f = 100$	$\sum fX = 6\,745$

$$\bar{X} = \frac{\sum fX}{\sum f} = \frac{\sum fX}{N} = \frac{6\,745}{100} = 67.45 \text{ in}$$

PROPIEDADES DE LA MEDIA ARITMÉTICA

3.16 Probar que la suma de las desviaciones de $X_1, X_2, \ldots, X_N$ respecto a su media $\bar{X}$ es igual a cero.

SOLUCIÓN

Sean $d_1 = X_1 - \bar{X}, d_2 = X_2 - \bar{X}, \ldots, d_N = X_N - \bar{X}$ las desviaciones de $X_1, X_2, \ldots, X_N$ de su media, $\bar{X}$. Entonces

$$\text{La suma de las desviaciones} = \sum d_j = \sum (X_j - \bar{X}) = \sum X_j - N\bar{X}$$

$$= \sum X_j - N\left(\frac{\sum X_j}{N}\right) = \sum X_j - \sum X_j = 0$$

donde se usa $\sum$ en vez de $\sum_{j=1}^{N}$. Si se desea, también se puede omitir el subíndice j de X_j siempre que éste se *sobreentienda*.

3.17 Si $Z_1 = X_1 + Y_1, Z_2 = X_2 + Y_2, \ldots, Z_N = X_N + Y_N$, probar que $\bar{Z} = \bar{X} + \bar{Y}$.

SOLUCIÓN

Por definición

$$\bar{X} = \frac{\sum X}{N} \qquad \bar{Y} = \frac{\sum Y}{N} \qquad \bar{Z} = \frac{\sum Z}{N}$$

Por lo tanto $\bar{Z} = \dfrac{\sum Z}{N} = \dfrac{\sum (X+Y)}{N} = \dfrac{\sum X + \sum Y}{N} = \dfrac{\sum X}{N} + \dfrac{\sum Y}{N} = \bar{X} + \bar{Y}$

en donde los subíndices de X, Y y Z se han omitido y donde $\sum$ significa $\sum_{j=1}^{N}$.

3.18 *a*) Si las desviaciones de N números $X_1, X_2, \ldots, X_N$ de un número cualquiera A están dadas por $d_1 = X_1 - A$, $d_2 = X_2 - A, \ldots, d_N = X_N - A$, respectivamente, probar que

$$\bar{X} = A + \frac{\sum_{j=1}^{N} d_j}{N} = A + \frac{\sum d}{N}$$

b) En caso de que $X_1, X_2, \ldots, X_K$ tengas frecuencias respectivas $f_1, f_2, \ldots, f_K$ y que $d_1 = X_1 - A, \ldots, d_K = X_K - A$ demostrar que en lugar del resultado del inciso *a*) se tiene

$$\bar{X} = A + \frac{\sum_{j=1}^{K} f_j d_j}{\sum_{j=1}^{K} f_j} = A + \frac{\sum fd}{N} \qquad \text{donde} \qquad \sum_{j=1}^{K} f_j = \sum f = N$$

SOLUCIÓN

a) **Primer método**

Ya que $d_j = X_j - A$ y que $X_j = A + d_j$, se tiene

$$\bar{X} = \frac{\sum X_j}{N} = \frac{\sum (A + d_j)}{N} = \frac{\sum A + \sum d_j}{N} = \frac{NA + \sum d_j}{N} = A + \frac{\sum d_j}{N}$$

donde se usa $\sum$ en lugar de $\sum_{j=1}^{N}$ para abreviar.

Segundo método

Se tiene $d = X - A$ o bien $X = A + d$, omitiendo los subíndices de d y X. Por lo tanto, de acuerdo con el problema 3.17,

$$\bar{X} = \bar{A} + \bar{d} = A + \frac{\sum d}{N}$$

ya que la media de cualquier cantidad de constantes todas iguales a A es A.

b)
$$\bar{X} = \frac{\sum_{j=1}^{K} f_j X_j}{\sum_{j=1}^{K} f_j} = \frac{\sum f_j X_j}{N} = \frac{\sum f_j (A + d_j)}{N} = \frac{\sum A f_j + \sum f_j d_j}{N} = \frac{A \sum f_j + \sum f_j d_j}{N}$$

$$= \frac{AN + \sum f_j d_j}{N} = A + \frac{\sum f_j d_j}{N} = A + \frac{\sum fd}{N}$$

Obsérvese que *formalmente* este resultado se obtiene del inciso *a*) sustituyendo d_j *por* $f_j d_j$ *y sumando desde* $j = 1$ hasta K en lugar de desde $j = 1$ hasta N. El resultado es equivalente a $\bar{X} = A + \bar{d}$, donde $\bar{d} = (\sum fd)/N$.

CÁLCULO DE LA MEDIA ARITMÉTICA A PARTIR DE DATOS AGRUPADOS

3.19 Emplee el método del problema 3.18*a*) para hallar la media aritmética de los números 5, 8, 11, 9, 12, 6, 14 y 10, eligiendo como "media supuesta" A los valores *a*) 9 y *b*) 20.

SOLUCIÓN

a) Las desviaciones de los números dados respecto al 9 son $-4, -1, 2, 0, 3, -3, 5$ y 1, y la suma de las desviaciones es $\sum d = -4 - 1 + 2 + 0 + 3 - 3 + 5 + 1 = 3$. Por lo tanto

$$\bar{X} = A + \frac{\sum d}{N} = 9 + \frac{3}{8} = 9.375$$

b) Las desviaciones de los números dados, respecto al 20, son $-15, -12, -9, -11, -8, -14, -6$ y -10 y $\sum d = -85$. Por lo tanto,

$$\bar{X} = A + \frac{\sum d}{N} = 20 + \frac{(-85)}{8} = 9.375$$

3.20 Emplee el método del problema 3.18*b*) para hallar la media aritmética de las estaturas de 100 estudiantes de la universidad XYZ (ver problema 3.15).

SOLUCIÓN

Para facilitar los cálculos pueden organizarse los datos como en la tabla 3.2. Como media supuesta A se toma la marca de clase 67 (que corresponde a la clase con mayor frecuencia), aunque para A se puede tomar cualquier marca de clase. Obsérvese que de esta manera los cálculos son más sencillos que en el problema 3.15. Para simplificar aún más el trabajo, se puede proceder como en el problema 3.22, donde se hace uso de que todas las desviaciones (columna 2 de la tabla 3.2) son múltiplos enteros de la amplitud del intervalo de clase.

Tabla 3.2

Marcas de clase (X)	Desviación $d = X - A$	Frecuencias (f)	fd
61	-6	5	-30
64	-3	18	-54
$A \rightarrow 67$	0	42	0
70	3	27	81
73	6	8	48
		$N = \sum f = 100$	$\sum fd = 45$

$$\bar{X} = A + \frac{\sum fd}{N} = 67 + \frac{45}{100} = 67.45 \text{ in}$$

3.21 Con $d_j = X_j - A$ se denotan las desviaciones de las marcas de clase X_j, de una distribución de frecuencias, respecto a una marca de clase dada A. Mostrar que si todos los intervalos de clase son de una misma amplitud c, entonces: *a*) todas las desviaciones son múltiplos de c (es decir, $d_j = cu_j$ donde $u_j = 0, \pm 1, \pm 2, \ldots$) y *b*) que la media aritmética se puede calcular empleando la fórmula

$$\bar{X} = A + \left(\frac{\sum fu}{N}\right)c$$

SOLUCIÓN

a) Lo pedido queda ilustrado en la tabla 3.2 del problema 3.20, donde en la columna 2 se observa que todas las desviaciones son múltiplos de la amplitud del intervalo de clase $c = 3$ in.

Para ver que esto es válido en general, obsérvese que si $X_1, X_2, X_3, \ldots$ son marcas de clase sucesivas, la diferencia entre ellas será igual a c, de manera que $X_2 = X_1 + c$, $X_3 = X_1 + 2c$, y en general, $X_j = X_1 + (j - 1)c$. Entonces, la diferencia entre cualesquiera dos marcas de clase, por ejemplo, X_p y X_q, será

$$X_p - X_q = [X_1 + (p - 1)c] - [X_1 + (q - 1)c] = (p - q)c$$

que es un múltiplo de c.

b) De acuerdo con el inciso *a*), las desviaciones de todas las marcas de clase respecto a una marca de clase dada son múltiplos de c (es decir, $d_j = cu_j$). Entonces, usando el problema 3.18*b*), se tiene

$$\bar{X} = A + \frac{\sum f_j d_j}{N} = A + \frac{\sum f_j (cu_j)}{N} = A + c\,\frac{\sum f_j u_j}{N} = A + \left(\frac{\sum fu}{N}\right)c$$

Obsérvese que esto es equivalente a $\bar{X} = A + c\bar{u}$, lo que se obtiene de $\bar{X} = A + \bar{d}$ sustituyendo $d = cu$ y observando que $\bar{d} = c\bar{u}$ (ver problema 3.18).

3.22 Emplee los resultados del problema 3.21*b*) para hallar la estatura media de los 100 estudiantes de la universidad XYZ (ver problema 3.20).

SOLUCIÓN

Para facilitar los cálculos pueden organizarse los datos como en la tabla 3.3. A este método de le llama *método de compilación* y se recomienda usarlo siempre que sea posible.

Tabla 3.3

	X	u	f	fu
	61	−2	5	−10
	64	−2	18	−18
$A \longrightarrow$	67	0	42	0
	70	1	27	27
	73	2	8	16
			$N = 100$	$\sum fu = 15$

$$\bar{X} = A + \left(\frac{\sum fu}{N}\right)c = 67 + \left(\frac{15}{100}\right)(3) = 67.45 \text{ in}$$

3.23 Calcule el salario medio semanal de los 65 empleados de la empresa P&R a partir de la distribución de frecuencias de la tabla 2.5, empleando: *a*) el método largo y *b*) el método codificado.

SOLUCIÓN

En las tablas 3.4 y 3.5 se dan las soluciones de *a*) y *b*), respectivamente.

Tabla 3.4

X	f	fX
\$255.00	8	\$2 040.00
265.00	10	2 650.00
275.00	16	4 400.00
285.00	14	3 990.00
295.00	10	2 950.00
305.00	5	1 525.00
315.00	2	630.00
	$N = 65$	$\sum fX = \$18\,185.00$

Tabla 3.5

	X	u	f	fX
	\$255.00	−2	8	−16
	265.00	−1	10	−10
$A \longrightarrow$	275.00	0	16	0
	285.00	1	14	14
	295.00	2	10	20
	305.00	3	5	15
	315.00	4	2	8
			$N = 65$	$\sum fu = 31$

Puede suponerse que estas tablas introducen un error, ya que en realidad las marcas de clase son \$254.995, \$264.995, etc., y no \$255.00, \$265.00, etc. Sin embargo, con las marcas de clase de la tabla 3.4, $\bar{X}$ resulta ser \$279.76 en lugar de \$279.77, lo que es una diferencia despreciable.

$$\bar{X} = \frac{\sum fX}{N} = \frac{\$18\ 185.00}{65} = \$279.77 \qquad \bar{X} = A + \left(\frac{\sum fu}{N}\right)c = \$275.00 + \frac{31}{65}(\$10.00) = \$279.77$$

3.24 Empleando la tabla 2.9*d*), hallar el salario medio de los 70 empleados de la empresa P&R.

SOLUCIÓN

En este caso, los intervalos de clase no son todos de la misma amplitud, por lo que se tiene que usar el método largo, como se muestra en la tabla 3.6

Tabla 3.6

X	u	fX
\$255.00	8	\$2 040.00
265.00	10	2 650.00
275.00	16	4 400.00
285.00	15	4 275.00
295.00	10	2 950.00
310.00	8	2 480.00
350.00	3	1 050.00
	$N = 70$	$\sum fX = \$19\ 845.00$

$$\bar{X} = \frac{\sum fX}{N} = \frac{\$19\ 845.00}{70} = \$283.50$$

LA MEDIANA

3.25 En los resultados de MINITAB, a continuación, se presenta el tiempo, por semana, que 30 usuarios de Internet pasaron haciendo búsquedas, así como la mediana de estos 30 tiempos. Verificar la mediana. ¿Se considera que este promedio es típico (representativo) de estos 30 tiempos? Compárense los resultados con los hallados en el problema 3.8.

```
MTB > print c1
```

Muestra de datos

```
tiempo
3   4   4   5   5   5   5   5   5   6
6   6   6   7   7   7   7   7   8   8
9  10  10  10  10  10  10  12  55  60
```

```
MTB > median c1
```

Mediana de columna

```
Median of time = 7.0000
```

SOLUCIÓN

Obsérvese que los dos valores de en medio son 7 y que la media de estos dos valores de en medio es 7. En el problema 3.8 se encontró que la media es 10.4 horas. La mediana es más típica (representativa) de estos tiempos que la media.

3.26 En los cajeros automáticos de cinco lugares de una ciudad grande, se registró la cantidad de transacciones por día. Los datos fueron 35, 49, 225, 50, 30, 65, 40, 55, 52, 76, 48, 325, 47, 32 y 60. Encontrar: *a*) la cantidad mediana de transacciones y *b*) la cantidad media de transacciones.

SOLUCIÓN

a) Los datos ordenados de menor a mayor son 30, 32, 35, 40, 47, 48, 49, 50, 52, 55, 60, 65, 76, 225 y 325. Como la cantidad de datos es un número non, sólo hay un valor de enmedio, 50, que es la mediana buscada.

b) La suma de los 15 valores es 1 189. La media es $1\,189/15 = 79.257$.

Obsérvese que a la mediana no le afectan los dos valores extremos 225 y 325, en tanto que a la media sí. En este caso, la mediana es un mejor indicador de la cantidad promedio de transacciones diarias en los cajeros automáticos.

3.27 Si en una ordenación se tienen: *a*) 85 y *b*) 150 números, ¿cómo se encuentra la mediana de estos números?

SOLUCIÓN

a) Como 85 es un número non, sólo hay un valor de en medio, habiendo 42 números mayores que él y 42 números menores que él. Por lo tanto, la mediana es el número que ocupa la posición 43 de la ordenación.

b) Como 150 es un número par, hay dos valores de en medio con 74 números menores que ellos y 74 números mayores que ellos. Los dos números de en medio son los números en las posiciones 75 y 76 de la ordenación; su media aritmética es la mediana buscada.

3.28 A partir de los datos del problema 2.8, encontrar el peso mediano de los 40 estudiantes de la universidad estatal empleando: *a*) la distribución de frecuencias dada en la tabla 2.7 (reproducida aquí como tabla 3.7) y *b*) los datos originales.

SOLUCIÓN

a) **Primer método** (empleando la interpolación)

Se supone que los pesos de la tabla 3.7 están distribuidos de manera continua. En ese caso, la mediana es un peso tal que la mitad del total de las frecuencias ($40/2 = 20$) quede por encima de él y la mitad del total de las frecuencias quede por debajo de él.

Tabla 3.7

Peso (lb)	Frecuencias
118-126	3
127-135	5
136-144	9
145-153	12
154-162	5
163-171	4
172-180	2
	Total 40

La suma de las tres primeras frecuencias de clase es 3 + 5 + 9 = 17. Por lo tanto, para dar la frecuencia 20, que es la buscada, se necesitan tres más de los 12 casos que pertenecen a la cuarta clase. Como el cuarto intervalo de clase, 145-153,

en realidad corresponde a los pesos desde 144.5 hasta 153.5, la mediana debe encontrarse a 3/12 entre 144.5 y 153.5, es decir, la mediana es

$$144.5 + \frac{3}{12}(153.5 - 144.5) = 144.5 + \frac{3}{12}(9) = 146.8 \text{ lb}$$

Segundo método (empleando la fórmula)

Como las sumas de las primeras tres clases y de las primeras cuatro clases son, respectivamente, $3 + 5 + 9 = 17$ y $3 + 5 + 9 + 12 = 29$, la mediana se encuentra en la cuarta clase, que es, por lo tanto, la clase mediana. Entonces.

$L_1 =$ frontera inferior de clase de la clase mediana $= 144.5$
$N =$ número de datos $= 40$
$(\sum f)_1 =$ suma de las frecuencias de todas las clases anteriores a la clase mediana $= 3 + 5 + 9 = 17$
$f_{\text{mediana}} =$ frecuencia de la clase mediana $= 12$
$c =$ amplitud del intervalo de la clase mediana $= 9$

y por lo tanto

$$\text{Mediana} = L_1 + \left(\frac{N/2 - (\sum f)_1}{f_{\text{mediana}}}\right)c = 144.5 + \left(\frac{40/2 - 17}{12}\right)(9) = 146.8 \text{ lb}$$

b) Dispuestos en una ordenación, los pesos originales son

119, 125, 126, 128, 132, 135, 135, 135, 136, 138, 138, 140, 140, 142, 142, 144, 144, 145, 145, 146
146, 147, 147, 148, 149, 150, 150, 152, 153, 154, 156, 157, 158, 161, 163, 164, 165, 168, 173, 176

La mediana es la media aritmética de los pesos en las posiciones 20 y 21 de esta ordenación y es igual a 146 lb.

3.29 En la figura 3-3 se muestra una representación de tallo y hoja que proporciona el número de muertes en accidentes de tránsito en 2005 relacionados con el alcohol en los 50 estados y Washington, D.C.

Representación de tallo y hoja: Muertes

Representación de tallo y hoja: Muertes N = 51
Leaf Unit = 10

14	**0**	**22334556667889**
23	**1**	**122255778**
(7)	**2**	**0334689**
21	**3**	**124679**
15	**4**	**22669**
10	**5**	**012448**
4	**6**	**3**
3	**7**	
3	**8**	
3	**9**	
3	**10**	
3	**11**	
3	**12**	
3	**13**	
3	**14**	**7**
2	**15**	**6**
1	**16**	
1	**17**	**1**

Figura 3-3 MINITAB, representación de tallo y hoja de las muertes en accidentes de tránsito relacionados con el alcohol.

Encontrar la media, la mediana y la moda de las muertes relacionadas con el alcohol dadas en la figura 3-3.

SOLUCIÓN

La cantidad de muertes va de 20 a 1 710. La distribución es bimodal. Las dos modas son 60 y 120. Ambas se presentan tres veces.

La clase `(7) 2 0334689 es la clase mediana`. Es decir, la mediana se encuentra en esta clase. La mediana es el dato de en medio o el dato que ocupa la posición 26 en la ordenación. El dato en la posición 24 es 200, el dato en la posición 25 es 230 y el dato en la posición 26 es 230. Por lo tanto, la mediana es 230.

La suma de estos 51 datos es 16 660 y la media es $16\,660/51 = 326.67$.

3.30 Encontrar el salario mediano de los 65 empleados de la empresa P&R (ver el problema 2.3).

SOLUCIÓN

En este caso, $N = 65$ y $N/2 = 32.5$. Como la suma de las primeras dos y de las primeras tres frecuencias de clase son $8 + 10 = 18$ y $8 + 10 + 16 = 34$, respectivamente, la clase mediana es la tercera clase. Usando la fórmula,

$$\text{Mediana} = L_1 + \left(\frac{N/2 - (\sum f)_1}{f_{\text{mediana}}}\right)c = \$269.995 + \left(\frac{32.5 - 18}{16}\right)(\$10.00) = \$279.06$$

LA MODA

3.31 Encontrar la media, la mediana y la moda de los conjuntos: *a*) 3, 5, 2, 6, 5, 9, 5, 2, 8, 6 y *b*) 51.6, 48.7, 50.3, 49.5, 48.9.

SOLUCIÓN

a) En una ordenación, los números son 2, 2, 3, 5, 5, 5, 6, 6, 8 y 9.

$$\text{Media} = \tfrac{1}{10}(2 + 2 + 3 + 5 + 5 + 5 + 6 + 6 + 8 + 9) = 5.1$$

$$\text{Mediana} = \text{media aritmética de los dos valores de en medio} = \tfrac{1}{2}(5 + 5) = 5$$

$$\text{Moda} = \text{número que se presenta con mayor frecuencia} = 5$$

b) En una ordenación, los números son 48.7, 48.9, 49.5 50.3 y 51.6.

$$\text{Media} = \tfrac{1}{5}(48.7 + 48.9 + 49.5 + 50.3 + 51.6) = 49.8$$

$$\text{Mediana} = \text{número de en medio} = 49.5$$

$$\text{Moda} = \text{número que se presenta con mayor frecuencia (no existe uno aquí)}$$

3.32 Supóngase que se desea hallar la moda de los datos de la figura 3-29. Se puede usar el procedimiento "frequencies" de SAS para obtener el resultado siguiente. Observando el resultado dado por el procedimiento FREQ (figura 3-4), ¿cuáles son las modas de la cantidad de muertes relacionadas con el alcohol?

Procedimiento FREQ

Muertes

Muertes	Frecuencias	Porcentaje	Frecuencias acumuladas	Porcentajes acumulados
20	2	3.92	2	3.92
30	2	3.92	4	7.84
40	1	1.96	5	9.80
50	2	3.92	7	13.73
60	3	5.88	10	19.61
70	1	1.96	11	21.57
80	2	3.92	13	25.49
90	1	1.96	14	27.45
110	1	1.96	15	29.41
120	3	5.88	18	35.29
150	2	3.92	20	39.22
170	2	3.92	22	43.14
180	1	1.96	23	45.10
200	1	1.96	24	47.06
230	2	3.92	26	50.98
240	1	1.96	27	52.94
260	1	1.96	28	54.90
280	1	1.96	29	56.86
290	1	1.96	30	58.82
310	1	1.96	31	60.78
320	1	1.96	32	62.75
340	1	1.96	33	64.71
360	1	1.96	34	66.67
370	1	1.96	35	68.63
390	1	1.96	36	70.59
420	2	3.92	38	74.51
460	2	3.92	40	78.43
490	1	1.96	41	80.39
500	1	1.96	42	82.35
510	1	1.96	43	84.31
520	1	1.96	44	86.27
540	2	3.92	46	90.20
580	1	1.96	47	92.16
630	1	1.96	48	94.12
1470	1	1.96	49	96.08
1560	1	1.96	50	98.04
1710	1	1.96	51	100.00

Figura 3-4 SAS, resultados del procedimiento FREQ para la cantidad de decesos relacionados con el alcohol.

SOLUCIÓN

Estos datos son bimodales y las modas son 60 y 120. Esto se encuentra al observar los resultados de SAS, donde se nota que la frecuencia, tanto de 60 como de 120, es 3, que es mayor que todas las demás frecuencias.

3.33 Algunos paquetes de software para estadística tienen rutinas para encontrar la moda, pero en los casos en los que los datos son multimodales, no dan todas las modas. En la figura 3-5 considerar el resultado que se obtiene con SPSS.

¿Qué hace SPSS cuando se le pide que encuentre las modas?

Muertes

		Frecuencias	Porcentaje	Porcentajes válidos	Porcentajes acumulados
Válido	20.00	2	3.9	3.9	3.9
	30.00	2	3.9	3.9	7.8
	40.00	1	2.0	2.0	9.8
	50.00	2	3.9	3.9	13.7
	60.00	3	5.9	5.9	19.6
	70.00	1	2.0	2.0	21.6
	80.00	2	3.9	3.9	25.5
	90.00	1	2.0	2.0	27.5
	110.00	1	2.0	2.0	29.4
	120.00	3	5.9	5.9	35.3
	150.00	2	3.9	3.9	39.2
	170.00	2	3.9	3.9	43.1
	180.00	1	2.0	2.0	45.1
	200.00	1	2.0	2.0	47.1
	230.00	2	3.9	3.9	51.0
	240.00	1	2.0	2.0	52.9
	260.00	1	2.0	2.0	54.9
	280.00	1	2.0	2.0	56.9
	290.00	1	2.0	2.0	58.8
	310.00	1	2.0	2.0	60.8
	320.00	1	2.0	2.0	62.7
	340.00	1	2.0	2.0	64.7
	360.00	1	2.0	2.0	66.7
	370.00	1	2.0	2.0	68.6
	390.00	1	2.0	2.0	70.6
	420.00	2	3.9	3.9	74.5
	460.00	2	3.9	3.9	78.4
	490.00	1	2.0	2.0	80.4
	500.00	1	2.0	2.0	82.4
	510.00	1	2.0	2.0	84.3
	520.00	1	2.0	2.0	86.3
	540.00	2	3.9	3.9	90.2
	580.00	1	2.0	2.0	92.2
	630.00	1	2.0	2.0	94.1
	1 470.00	1	2.0	2.0	96.1
	1 560.00	1	2.0	2.0	98.0
	1 710.00	1	2.0	2.0	100.0
	Total	51	100.0	100.0	

Estadística

Muertes

N	Válido	51
	Equivocado	0
Moda		60.00[a]

[a] Hay múltiples modas. Se muestra el valor más pequeño.

Figura 3-5 SPSS, resultado para las muertes relacionadas con el alcohol.

SOLUCIÓN

SPSS da la moda más pequeña. Pero se puede inspeccionar la distribución de frecuencias y hallar las modas de la misma manera que con SAS (ver el resultado dado antes).

RELACIÓN EMPÍRICA ENTRE LA MEDIA, LA MEDIANA Y LA MODA

3.34 *a*) Emplear la fórmula empírica media − moda = 3(media − mediana) para hallar el salario modal de los 65 empleados de la empresa P&R.

b) Comparar el resultado con la moda obtenida en el problema 3.33.

SOLUCIÓN

a) De acuerdo con los problemas 3.23 y 3.30 se tiene media = \$279.77 y mediana = \$279.06. Por lo tanto,

$$\text{Moda} = \text{media} - 3(\text{media} - \text{mediana}) = \$279.77 - 3(\$279.77 - \$279.06) = \$277.64$$

b) De acuerdo con el problema 3.33, el salario modal es \$277.50, de manera que en este caso coincide con el resultado empírico.

LA MEDIA GEOMÉTRICA

3.35 Encontrar: *a*) la media geométrica y *b*) la media aritmética de los números 3, 5, 6, 6, 7, 10 y 12. Se supone que los números son exactos.

SOLUCIÓN

a) Media geométrica $= G = \sqrt[7]{(3)(5)(6)(6)(7)(10)(12)} = \sqrt[7]{453\,600}$. Empleando logaritmos comunes, $\log G = \frac{1}{7}\log 453\,600 = \frac{1}{7}(5.6567) = 0.8081$ y $G = 6.43$ (a la centésima más cercana). Otra posibilidad es usar una calculadora.

Otro método

$$\begin{aligned}\log G &= \tfrac{1}{7}(\log 3 + \log 5 + \log 6 + \log 6 + \log 7 + \log 10 + \log 12)\\ &= \tfrac{1}{7}(0.4771 + 0.6990 + 0.7782 + 0.7782 + 0.8451 + 1.0000 + 1.0792)\\ &= 0.8081\end{aligned}$$

y $G = 6.43$

b) Media aritmética $= \hat{X} = \frac{1}{7}(3 + 5 + 6 + 6 + 7 + 10 + 12) = 7$. Esto ilustra que la media geométrica de un conjunto de números positivos, no todos iguales, es menor que su media aritmética.

3.36 Los números $X_1, X_2, \ldots, X_K$ se presentan con frecuencias $f_1, f_2, \ldots, f_K$ donde $f_1 + f_2 + \ldots, + f_K = N$ es la frecuencia total.

a) Encontrar la media geométrica G de estos números.

b) Deducir una expresión para $\log G$.

c) ¿Cómo se pueden emplear los resultados para hallar la media geométrica de datos agrupados en una distribución de frecuencias?

SOLUCIÓN

a)
$$G = \sqrt[N]{\underbrace{X_1X_1\cdots X_1}_{f_1 \text{ veces}}\ \underbrace{X_2X_2\cdots X_2}_{f_2 \text{ veces}}\cdots\underbrace{X_KX_K\cdots X_K}_{f_K \text{ veces}}} = \sqrt[N]{X_1^{f_1}X_2^{f_2}\cdots X_K^{f_K}}$$

donde $N = \sum f$. A esta media suele llamársele *media geométrica ponderada.*

b)
$$\log G = \frac{1}{N}\log\left(X_1^{f_1}X_2^{f_2}\cdots X_K^{f_K}\right) = \frac{1}{N}\left(f_1\log X_1 + f_2\log X_2 + \cdots + f_K\log X_K\right)$$
$$= \frac{1}{N}\sum_{j=1}^{K} f_j \log X_j = \frac{\sum f \log X}{N}$$

donde se supone que todos los números son positivos; de otra manera, los logaritmos no están definidos.

Obsérvese que el logaritmo de una media geométrica de un conjunto de números positivos es la media aritmética de los logaritmos de los números.

c) Al hallar la media geométrica de datos agrupados, este resultado puede emplearse tomando $X_1, X_2, \ldots, X_K$ como las marcas de clase y $f_1, f_2, \ldots, f_K$ como sus frecuencias correspondientes.

3.37 Durante un año la relación entre precios de un cuarto de galón de leche respecto a precios de una barra de pan fue 3.00, en tanto que al año siguiente la relación fue 2.00.

a) Encontrar la media aritmética de esta relación en estos dos años.
b) Encontrar la media aritmética de las relaciones ahora entre los precios de una barra de pan respecto a los precios de un cuarto de galón de leche en este periodo de 2 años.
c) Analizar la conveniencia de emplear la media aritmética para promediar relaciones.
d) Analizar la idoneidad de la media geométrica para promediar relaciones.

SOLUCIÓN

a) Media de las relaciones (cocientes) precio de leche respecto a precios de pan $= \frac{1}{2}(3.00 + 2.00) = 2.50$.

b) Como el primer año la relación entre precios de leche respecto a precios de pan es 3.00, la relación entre precios de pan respecto a precios de leche es $1/3 = 0.333$. De igual manera, la relación entre precios de pan y precios de leche el segundo año es $1/2.00 = 0.500$.

Por lo tanto,

Media de las relaciones (cocientes) precio de pan respecto a precios de leche $= \frac{1}{2}(0.333 + 0.500) = 0.417$

c) Si la media fuera un promedio adecuado, se esperaría que la media de las relaciones de precios de leche respecto a precios de pan fuera el recíproco de la media de las relaciones precios de pan respecto a precios de leche. Sin embargo, $1/0.417 = 2.40 \neq 2.50$. Esto demuestra que la media no es un promedio adecuado para (cocientes) relaciones.

d) La media geométrica de las relaciones entre precios de leche respecto a precios de pan $= \sqrt{(3.00)(2.00)} = \sqrt{6.00}$

La media geométrica de las relaciones entres precios de pan respecto a precios de leche $= \sqrt{(0.333)(0.500)} = \sqrt{0.0167} = 1/\sqrt{6.00}$

Dado que estos promedios son recíprocos, se concluye que la media geométrica es más adecuada que la media aritmética para promediar relaciones (cocientes).

3.38 La cuenta bacteriana en cierto medio de cultivo aumentó de 1 000 a 4 000 en 3 días. ¿Cuál es el incremento porcentual promedio por día?

SOLUCIÓN

Como un incremento de 1 000 a 4 000 es un incremento de 300%, uno está inclinado a concluir que el aumento porcentual promedio por día es $300\%/3 = 100\%$. Sin embargo, esto significaría que el primer día la cuenta aumentó de 1 000 a 2 000, el segundo día de 2 000 a 4 000 y el tercer día de 4 000 a 8 000, lo cual no es así.

Para determinar este incremento porcentual promedio se denotará r a este incremento porcentual promedio. Entonces

$$\text{Cuenta bacteriana total un día después} = 1\,000 + 1\,000r = 1\,000(1 + r)$$
$$\text{Cuenta bacteriana total dos días después} = 1\,000(1 + r) + 1\,000(1 + r)r = 1\,000(1 + r)^2$$
$$\text{Cuenta bacteriana total tres días después} = 1\,000(1 + r)^2 + 1\,000(1 + r)^2 r = 1\,000(1 + r)^3$$

Esta última expresión debe ser igual a 4 000. De manera que $1\,000(1 + r)^3 = 4\,000$, $(1 + r)^3 = 4$, $1 + r = \sqrt[3]{4}$, y $r = \sqrt[3]{4} - 1 = 1.587 - 1 = 0.587$, y así, $r = 58.7\%$.

En general, si se parte de una cantidad P y se incrementa esta cantidad a una tasa constante r por unidad de tiempo, la cantidad que se tendrá después de n unidades de tiempo será

$$A = P(1 + r)^n$$

A esta fórmula se le llama *fórmula del interés compuesto* (ver problemas 3.94 y 3.95).

LA MEDIA ARMÓNICA

3.39 Encontrar la media armónica H de los números 3, 5, 6, 6, 7, 10 y 12.

SOLUCIÓN

$$\frac{1}{H} = \frac{1}{N}\sum\frac{1}{X} = \frac{1}{7}\left(\frac{1}{3}+\frac{1}{5}+\frac{1}{6}+\frac{1}{6}+\frac{1}{7}+\frac{1}{10}+\frac{1}{12}\right) = \frac{1}{7}\left(\frac{140+84+70+70+60+42+35}{420}\right)$$

$$= \frac{501}{2\,940}$$

y $$H = \frac{2\,940}{501} = 5.87$$

Suele ser mejor expresar primero las fracciones en forma decimal. Así

$$\frac{1}{H} = \tfrac{1}{7}(0.3333 + 0.2000 + 0.1667 + 0.1667 + 0.1429 + 0.1000 + 0.0833)$$

$$= \frac{1.1929}{7}$$

y $$H = \frac{7}{1.1929} = 5.87$$

Comparando con los resultados del problema 3.35 se ilustra el hecho de que la media armónica de números positivos, no todos iguales, es menor que su media geométrica, la que a su vez es menor que su media aritmética.

3.40 Durante cuatro años consecutivos los precios del fuel para la calefacción son \$0.80, \$0.90, \$1.05 y \$1.25 por galón (gal). ¿Cuál es el precio promedio del fuel en estos cuatro años?

SOLUCIÓN

Caso 1

Supóngase que todos los años se compra la misma cantidad de fuel, digamos 1 000 gal. Entonces

$$\text{Precio promedio} = \frac{\text{precio total}}{\text{cantidad total comprada}} = \frac{\$800 + \$900 + \$1\,050 + \$1\,250}{4\,000\,\text{gal}} = \$1.00/\text{gal}$$

Esto es lo mismo que la media aritmética del costo por galón; es decir $\frac{1}{4}(\$0.80 + \$0.90 + \$1.05 + \$1.25) = 1.00/\text{gal}$. Este resultado sería el mismo aun cuando se usaran x galones por año.

Caso 2

Supóngase que en el fuel se gasta la misma cantidad de dinero todos los años, o sea \$1 000. Entonces

$$\text{Precio promedio} = \frac{\text{precio total}}{\text{cantidad total comprada}} = \frac{\$4\,000}{(1\,250 + 1\,111 + 952 + 800)\text{gal}} = \$0.975/\text{gal}$$

Esto es lo mismo que la media armónica de los precios por galón:

$$\frac{4}{\frac{1}{0.80}+\frac{1}{0.90}+\frac{1}{1.05}+\frac{1}{1.25}} = 0.975$$

El resultado será el mismo si se gastan y dólares por año.

Ambos promedios son correctos, pero se calculan para condiciones diferentes.

Debe notarse que si la cantidad de galones empleados varía de un año a otro, en vez de ser siempre la misma, en lugar de la media aritmética ordinaria usada en el caso 1, hay que usar la media aritmética ponderada. De manera similar, si la cantidad gastada varía de un año a otro, en lugar de la media armónica empleada en el caso 2 se debe usar la media armónica ponderada.

3.41 Un automóvil recorre 25 millas a 25 millas por hora (mph), 25 millas a 50 mph y 25 millas a 75 mph. Encontrar la media aritmética de las tres velocidades y la media armónica de las tres velocidades. ¿Cuál es correcta?

SOLUCIÓN

La velocidad promedio es igual a la distancia recorrida dividida entre el total del tiempo y es igual a lo siguiente:

$$\frac{75}{\left(1+\frac{1}{2}+\frac{1}{3}\right)} = 40.9 \text{ mi/h}$$

La media aritmética de las tres velocidades es:

$$\frac{25+50+75}{3} = 50 \text{ mi/h}$$

La media armónica se encuentra como sigue:

$$\frac{1}{H} = \frac{1}{N}\sum\frac{1}{X} = \frac{1}{3}\left(\frac{1}{25}+\frac{1}{50}+\frac{1}{75}\right) = \frac{11}{450} \qquad \text{y} \qquad H = \frac{450}{11} = 40.9$$

La media armónica es la medida correcta de la velocidad promedio.

LA RAÍZ CUADRADA MEDIA O MEDIA CUADRÁTICA

3.42 Encontrar la media cuadrática de los números 3, 5, 6, 6, 7, 10 y 12.

SOLUCIÓN

$$\text{Media cuadrática} = \text{RCM} = \sqrt{\frac{3^2+5^2+6^2+6^2+7^2+10^2+12^2}{7}} = \sqrt{57} = 7.55$$

3.43 Demostrar que la media cuadrática de dos números positivos distintos a y b es mayor que su media geométrica.

SOLUCIÓN

Se pide que se demuestre que $\sqrt{\frac{1}{2}(a^2+b^2)} > \sqrt{ab}$. Si esto es verdad, entonces elevando al cuadrado ambos miembros $\frac{1}{2}(a^2+b^2) > ab$, de manera que $a^2+b^2 > 2ab$, $a^2-2ab^2+b^2 > 0$ o bien $(a-b)^2 > 0$. Pero esta igualdad es cierta, ya que el cuadrado de cualquier número real distinto de cero es positivo.

La prueba consiste en demostrar el proceso inverso. Entonces, partiendo de $(a-b)^2 > 0$, que se sabe que es verdadero, se puede mostrar que $a^2 + b^2 > 2ab$, $\frac{1}{2}(a^2 + b^2) > ab$ y finalmente $\sqrt{\frac{1}{2}(a^2 + b^2)} > \sqrt{ab}$, que es lo pedido.

Obsérvese que $\sqrt{\frac{1}{2}(a^2 + b^2)} = \sqrt{ab}$ si y sólo si $a = b$.

CUARTILES, DECILES Y PERCENTILES

3.44 Para los salarios de los 65 empleados de la empresa P&R (ver problema 2.9), encontrar: *a*) los cuartiles Q_1, Q_2 y Q_3 y *b*) los deciles D_1, $D_2, \ldots, D_9$.

SOLUCIÓN

a) El primer cuartil Q_1 es el salario que se encuentra contando $N/4 = 65/4 = 16.25$ de los casos, comenzando con la primera clase (la más baja). Como la primera clase contiene 8 casos, hay que tomar 8.5 (16.25 − 8) casos de los 10 de la segunda clase. Usando el método de interpolación lineal, se tiene

$$Q_1 = \$259.995 + \frac{8.25}{10}(\$10.00) = \$268.25$$

El segundo cuartil Q_2 se encuentra contando los primeros $2N/4 = N/2 = 65/2 = 32.5$ de los casos. Como las primeras dos clases comprenden 18 casos, se deben tomar 32.5 – 18 = 14.5 casos de los 16 de la tercera clase, por lo tanto

$$Q_2 = \$269.995 + \frac{14.5}{16}(\$10.00) = \$279.06$$

Obsérvese que Q_2 es la mediana.

El tercer cuartil Q_3 se encuentra contando los primeros $3N/4 = \frac{3}{4}(65) = 48.75$ de los casos. Como las primeras cuatro clases comprenden 48 casos, se deben tomar 48.75 – 48 = 0.75 casos de los 10 de la quinta clase; por lo tanto

$$Q_3 = \$289.995 + \frac{0.75}{10}(\$10.00) = \$290.75$$

Así, 25% de los empleados ganan \$268.25 o menos, 50% gana \$279.06 o menos y 75% gana \$290.75 o menos.

b) Los deciles primero, segundo, ..., y noveno se obtienen contando $N/10$, $2N/10, \ldots, 9N/10$ de los casos empezando por la primer clase (inferior). Por lo tanto

$$D_1 = \$249.995 + \frac{6.5}{8}(\$10.00) = \$258.12 \qquad D_6 = \$279.995 + \frac{5}{14}(\$10.00) = \$283.57$$

$$D_2 = \$259.995 + \frac{5}{10}(\$10.00) = \$265.00 \qquad D_7 = \$279.995 + \frac{11.5}{14}(\$10.00) = \$288.21$$

$$D_3 = \$269.995 + \frac{1.5}{16}(\$10.00) = \$270.94 \qquad D_8 = \$289.995 + \frac{4}{10}(\$10.00) = \$294.00$$

$$D_4 = \$269.995 + \frac{8}{16}(\$10.00) = \$275.00 \qquad D_9 = \$299.995 + \frac{0.5}{5}(\$10.00) = \$301.00$$

$$D_5 = \$269.995 + \frac{14.5}{16}(\$10.00) = \$279.06$$

De manera que 10% de los empleados gana \$258.12 o menos, 20% gana \$265.00 o menos, ..., 90% gana \$301.00 o menos.

Obsérvese que el quinto decil es la mediana. Los deciles segundo, cuarto, sexto y octavo, que dividen la distribución en cinco partes iguales y a los que se les llama *quintiles,* también suelen usarse en la práctica.

3.45 En la distribución del problema 3.44, determinar *a*) el percentil 35o. y *b*) el percentil 60o.

SOLUCIÓN

a) El percentil 35o., que se denota P_{35}, se obtiene contando los primeros $35N/100 = 35(65)/100 = 22.75$ casos, empezando en la primera clase (la clase más baja). Entonces, como en el problema 3.44,

$$P_{35} = \$269.995 + \frac{4.75}{16}(\$10.00) = \$272.97$$

Esto significa que 35% de los empleados gana \$272.97 o menos.

b) El percentil 60o. es $P_{60} = \$279.995 + \frac{5}{14}(\$10.00) = \$283.57$. Obsérvese que éste coincide con el sexto decil o tercer quintil.

3.46 La siguiente hoja de cálculo de EXCEL está contenida en A1:D26. Esta hoja de cálculo contiene el ingreso per cápita en cada uno de los 50 estados de Estados Unidos. Dar los comandos de EXCEL para hallar Q_1, Q_2, Q_3 y P_{95}. Dar también los estados que están a ambos lados de estos cuartiles o percentiles.

Estado	Ingreso per cápita	Estado	Ingreso per cápita
Wyoming	36 778	Pennsylvania	34 897
Montana	29 387	Wisconsin	33 565
North Dakota	31 395	Massachusetts	44 289
New Mexico	27 664	Missouri	31 899
West Virginia	27 215	Idaho	28 158
Rhode Island	36 153	Kentucky	28 513
Virginia	38 390	Minnesota	37 373
South Dakota	31 614	Florida	33 219
Alabama	29 136	South Carolina	28 352
Arkansas	26 874	New York	40 507
Maryland	41 760	Indiana	31 276
Iowa	32 315	Connecticut	47 819
Nebraska	33 616	Ohio	32 478
Hawaii	34 539	New Hampshire	38 408
Mississippi	25 318	Texas	32 462
Vermont	33 327	Oregon	32 103
Maine	31 252	New Jersey	43 771
Oklahoma	29 330	California	37 036
Delaware	37 065	Colorado	37 946
Alaska	35 612	North Carolina	30 553
Tennessee	31 107	Illinois	36 120
Kansas	32 836	Michigan	33 116
Arizona	30 267	Washington	35 409
Nevada	35 883	Georgia	31 121
Utah	28 061	Louisiana	24 820

SOLUCIÓN

		Estados más cercanos
=PERCENTILE(A2:D26,0.25)	\$30 338.5	Arizona y NorthCarolina
=PERCENTILE(A2:D26,0.50)	\$32 657	Ohio y Kansas
=PERCENTILE(A2:D26,0.75)	\$36 144.75	Illinois y RhodeIsland
=PERCENTILE(A2:D26,0.95)	\$42 866.05	Maryland y NewJersey

PROBLEMAS SUPLEMENTARIOS

SUMATORIA

3.47 Escribir los términos de cada una de las sumas siguientes:

a) $\sum_{j=1}^{4}(X_j+2)$ *c)* $\sum_{j=1}^{3}U_j(U_j+6)$ *e)* $\sum_{j=1}^{4}4X_jY_j$

b) $\sum_{j=1}^{5}f_jX_j^2$ *d)* $\sum_{k=1}^{N}(Y_k^2-4)$

3.48 Escribir cada una de las sumas siguientes usando el signo de sumatoria:

a) $(X_1+3)^3+(X_2+3)^3+(X_3+3)^3$

b) $f_1(Y_1-a)^2+f_2(Y_2-a)^2+\cdots+f_{15}(Y_{15}-a)^2$

c) $(2X_1-3Y_1)+(2X_2-3Y_2)+\cdots+(2X_N-3Y_N)$

d) $(X_1/Y_1-1)^2+(X_2/Y_2-1)^2+\cdots+(X_8/Y_8-1)^2$

e) $\dfrac{f_1a_1^2+f_2a_2^2+\cdots+f_{12}a_{12}^2}{f_1+f_2+\cdots+f_{12}}$

3.49 Demostrar que $\sum_{j=1}^{N}(X_j-1)^2=\sum_{j=1}^{N}X_j^2-2\sum_{j=1}^{N}X_j+N$

3.50 Demostrar que $\sum(X+a)(Y+b)=\sum XY+a\sum Y+b\sum X+Nab$, donde a y b son constantes. ¿Cuáles son los subíndices implícitos?

3.51 Las variables U y V toman los valores $U_1=3$, $U_2=-2$, $U_3=5$ y $V_1=-4$, $V_2=-1$, $V_3=6$, respectivamente. Calcular *a)* $\sum UV$, *b)* $\sum(U+3)(V-4)$, *c)* $\sum V^2$, *d)* $(\sum U)(\sum V)^2$, *e)* $\sum UV^2$, *f)* $\sum(U^2-2V^2+2)$ y *g)* $\sum(U/V)$.

3.52 Dado que $\sum_{j=1}^{4}X_j=7$, $\sum_{j=1}^{4}Y_j=-3$ y $\sum_{j=1}^{4}X_jY_j=5$, encontrar *a)* $\sum_{j=1}^{4}(2X_j+5Y_j)$ y *b)* $\sum_{j=1}^{4}(X_j-3)(2Y_j+1)$.

LA MEDIA ARITMÉTICA

3.53 En cinco materias, un estudiante obtuvo las calificaciones siguientes: 85, 76, 93, 82 y 96. Determinar la media aritmética de estas calificaciones.

3.54 Un psicólogo mide los tiempos de reacción de un individuo a ciertos estímulos; éstos fueron 0.53, 0.46, 0.50, 0.49,0.52, 0.53, 0.44 y 0.55 segundos, respectivamente. Estimar el tiempo medio de reacción del individuo a estos estímulos.

3.55 Un conjunto de números consta de 6 seises, 7 sietes, 8 ochos, 9 nueves y 10 dieces. ¿Cuál es la media aritmética de estos números?

3.56 Un estudiante obtuvo las calificaciones siguientes en tres aspectos de un curso: 71, 78 y 89, respectivamente.

a) Si los pesos que se acuerda dar a estas calificaciones son 2, 4 y 5, respectivamente, ¿cuál es una calificación promedio apropiada?

b) ¿Cuál es la calificación promedio si se usan pesos iguales?

3.57 Los promedios de calificación en los cursos de tres maestros de economía son 79, 74 y 82, y sus grupos constan de 32, 25 y 17 alumnos, respectivamente. Determinar la calificación media de los tres cursos.

3.58 El salario anual medio pagado a los empleados de una empresa es \$36 000. Los salarios anuales medios pagados a hombres y mujeres de la empresa son \$34 000 y \$40 000, respectivamente. Determinar el porcentaje de hombres y mujeres empleados por la empresa.

3.59 En la tabla 3.8 se presenta la distribución de las cargas máximas, en toneladas cortas (1 tonelada corta = 2 000 lb) que soportan ciertos cables producidos por una empresa. Determinar la carga máxima media usando: *a*) el método largo y *b*) el método de compilación.

Tabla 3.8

Carga máxima (toneladas cortas)	Cantidad de cables
9.3-9.7	2
9.8-10.2	5
10.3-10.7	12
10.8-11.2	17
11.3-11.7	14
11.8-12.2	6
12.3-12.7	3
12.8-13.2	1
	Total 60

3.60 Encontrar $\bar{X}$ para los datos de la tabla 3.9 usando: *a*) el método largo y *b*) el método de compilación.

Tabla 3.9

X	462	480	498	516	534	552	570	588	606	624
f	98	75	56	42	30	21	15	11	6	2

3.61 En la tabla 3.10 se presenta la distribución de los diámetros de las cabezas de remaches producidos por una empresa. Calcular el diámetro medio.

3.62 Calcular la media de los datos de la tabla 3.11.

Tabla 3.10

Diámetro (cm)	Frecuencias
0.7247-0.7249	2
0.7250-0.7252	6
0.7253-0.7255	8
0.7256-0.7258	15
0.7259-0.7261	42
0.7262-0.7264	68
0.7265-0.7267	49
0.7268-0.7270	25
0.7271-0.7273	18
0.7274-0.7276	12
0.7277-0.7279	4
0.7280-0.7282	1
	Total 250

Tabla 3.11

Clase	Frecuencias
10 hasta menos de 15	3
15 hasta menos de 20	7
20 hasta menos de 25	16
25 hasta menos de 30	12
30 hasta menos de 35	9
35 hasta menos de 40	5
40 hasta menos de 45	2
	Total 54

3.63 Calcular la media de la cantidad de tiempo que ven televisión los 400 estudiantes del problema 2.20.

3.64 *a*) Emplear la distribución de frecuencias del problema 2.27 para calcular el diámetro medio de los balines.
b) Calcular la media directamente de los datos en bruto y compararla con el inciso *a*); explicar cualquier discrepancia.

LA MEDIANA

3.65 Encontrar la media y la mediana de estos conjuntos de números: *a*) 5, 4, 8, 3, 7, 2, 9 y *b*) 18.3, 20.6, 19.3, 22.4, 20.2, 18.8, 19.7, 20.0.

3.66 Encontrar la calificación mediana del problema 3.53.

3.67 Encontrar el tiempo mediano de reacción del problema 3.54.

3.68 Encontrar la mediana del conjunto de números del problema 3.55.

3.69 Encontrar la mediana de la carga máxima de los cables de la tabla 3.8 del problema 3.59.

3.70 Encontrar la mediana $\tilde{X}$ de la distribución presentada en la tabla 3.9 del problema 3.60.

3.71 Encontrar el diámetro mediano de las cabezas de los remaches de la tabla 3.10 del problema 3.61.

3.72 Encontrar la mediana de la distribución presentada en la tabla 3.11 del problema 3.62.

3.73 En la tabla 3.12 se da la cantidad, en miles, de muertes en Estados Unidos ocurridas en 1993 a causa de enfermedades cardiacas. Encontrar la edad mediana.

Tabla 3.12

Grupo de edad	Miles de muertes
Total	743.3
Menos de 1	0.7
1 a 4	0.3
5 a 14	0.3
15 a 24	1.0
25 a 34	3.5
35 a 34	13.1
45 a 54	32.7
55 a 64	72.0
65 a 74	158.1
75 a 84	234.0
85 y más	227.6

Fuente: U.S. National Center for Health Statistics, Vital Statistics of the U.S., annual.

3.74 Con los datos de la tabla del problema 2.31 encontrar la edad mediana.

3.75 Encontrar la mediana de la cantidad de tiempo que ven la televisión los 400 estudiantes del problema 2.20.

LA MODA

3.76 Encontrar la media, la mediana y la moda de cada uno de los conjuntos de números siguientes: *a*) 7, 4, 10, 9, 15, 12, 7, 9, 7 y *b*) 8, 11, 4, 3, 2, 5, 10, 6, 4, 1, 10, 8, 12, 6, 5, 7.

3.77 En el problema 3.53 encontrar la calificación modal.

3.78 En el problema 3.54 encontrar el tiempo de reacción modal.

3.79 En el problema 3.55 encontrar la moda del conjunto de números.

3.80 En el problema 3.59 encontrar la moda de la carga máxima de los cables.

3.81 En el problema 3.60 encontrar la moda $\hat{X}$ de la distribución dada en la tabla 3.9.

3.82 En el problema 3.61 encontrar el diámetro modal de las cabezas de los remaches de la tabla 3.10.

3.83 En el problema 3.62 encontrar la moda de la distribución dada.

3.84 En el problema 2.20 encontrar la moda de la cantidad de tiempo que ven televisión los 400 estudiantes.

3.85 *a*) ¿Cuál es el grupo de edad modal en la tabla 2.15?
b) ¿Cuál es el grupo de edad modal en la tabla 3.12?

3.86 Empleando las fórmulas (*9*) y (*10*) de este capítulo, hallar la moda de las distribuciones dadas en los problemas siguientes. Comparar las respuestas obtenidas con cada una de las dos fórmulas.
a) Problema 3.59 *b*) Problema 3.61 *c*) Problema 3.62 *d*) Problema 2.20.

3.87 La probabilidad de una variable aleatoria continua está descrita por la siguiente función de densidad de probabilidad. $f(x) = -0.75x^2 + 1.5x$ para $0 < x < 2$ y para todos los demás valores de x, $f(x) = 0$. La moda se presenta en el punto en el que la función alcanza su máximo. Empleando los conocimientos sobre funciones cuadráticas, mostrar que la moda se presenta en $x = 1$.

LA MEDIA GEOMÉTRICA

3.88 Hallar la media geométrica de los números: *a*) 4.2 y 16.8 y *b*) 3.00 y 6.00.

3.89 Hallar: *a*) la media geométrica G y *b*) la media aritmética $\bar{X}$ del conjunto 2, 4, 8, 16, 32.

3.90 Hallar la media geométrica de los conjuntos: *a*) 3, 5, 8, 3, 7, 2 y *b*) 28.5, 73.6, 47.2, 31.5, 64.8.

3.91 Hallar la media geométrica de las distribuciones de: *a*) el problema 3.59 y *b*) el problema 3.60. Verificar que en estos casos la media geométrica es menor o igual a la media aritmética.

3.92 Si en un periodo de 4 años se duplican los precios de un artículo, ¿cuál es el incremento porcentual anual promedio?

3.93 En 1980 y 1996 la población de Estados Unidos era de 226.5 millones y 266.0 millones, respectivamente. Empleando la fórmula dada en el problema 3.38, contestar lo siguiente.

a) ¿Cuál es el incremento porcentual anual promedio?

b) Estimar la población en 1985.

c) Si el incremento porcentual anual promedio de 1996 a 2000 es el mismo que en el inciso *a*), ¿a cuánto ascenderá la población en 2000?

3.94 Se invierten \$1 000 a una tasa de interés anual de 8%. ¿A cuánto ascenderá la cantidad total después de 6 años si no se retira el capital inicial?

3.95 Si en el problema 3.94 el interés es compuesto trimestralmente (es decir, el dinero gana 2% de interés cada 3 meses), ¿cuál será la cantidad total después de 6 años?

3.96 Encontrar dos números cuya media aritmética sea 9.00 y cuya media geométrica sea 7.2.

LA MEDIA ARMÓNICA

3.97 Encontrar la media armónica de los números: *a*) 2, 3 y 6 y *b*) 3.2, 5.2, 4.8, 6.1 y 4.2.

3.98 Encontrar: *a*) la media aritmética, *b*) la media geométrica y *c*) la media armónica de los números 0, 2, 4 y 6.

3.99 Si $X_1, X_2, X_3, \ldots$, son las marcas de clase de una distribución de frecuencias y $f_1, f_2, f_3, \ldots$, son sus frecuencias correspondientes, demostrar que su media armónica está dada por

$$\frac{1}{H} = \frac{1}{N}\left(\frac{f_1}{X_1} + \frac{f_2}{X_2} + \frac{f_3}{X_3} + \cdots\right) = \frac{1}{N}\sum \frac{f}{X}$$

donde $N = f_1 + f_2 + \cdots = \sum f$

3.100 Emplear el problema 3.99 para hallar la media armónica de la distribución: *a*) del problema 3.59 y *b*) del problema 3.60. Comparar con el problema 3.91.

3.101 Las ciudades *A*, *B* y *C* están equidistantes una de otra. Un conductor viaja de la ciudad *A* a la ciudad *B* a 30 mi/h, de la ciudad *B* a la ciudad *C* a 40 mi/h y de la ciudad *C* a la ciudad *A* a 50 mi/h. Determinar su velocidad promedio en este viaje.

3.102 *a*) Un aeroplano recorre las distancias d_1, d_2 y d_3 a las velocidades v_1, v_2 y v_3 mi/h, respectivamente. Mostrar que la velocidad promedio está dada por V, donde

$$\frac{d_1 + d_2 + d_3}{V} = \frac{d_1}{v_1} + \frac{d_2}{v_2} + \frac{d_3}{v_3}$$

Ésta es una media armónica ponderada.

b) Encontrar: V si $d_1 = 2\,500$, $d_2 = 1\,200$, $d_3 = 500$, $v_1 = 500$, $v_2 = 400$ y $v_3 = 250$.

3.103 Demostrar que la media geométrica de dos números *a* y *b* es: *a*) menor o igual que su media aritmética y *b*) mayor o igual que su media armónica. ¿Puede generalizar la prueba a más de dos números?

LA RAÍZ CUADRADA MEDIA O LA MEDIA CUADRÁTICA

3.104 Encontrar la RCM (o media cuadrática) de los números: *a*) 11, 23 y 35, y *b*) 2.7, 3.8, 3.2 y 4.3.

3.105 Probar que la RCM de dos números positivos, *a* y *b*, es: *a*) mayor o igual que la media aritmética y *b*) mayor o igual que la media armónica. Se puede extender la prueba a más de dos números.

3.106 Deducir una fórmula que pueda usarse para hallar la RCM de datos agrupados y aplicarla a una de las distribuciones de frecuencias ya consideradas.

CUARTILES, DECILES Y PERCENTILES

3.107 En la tabla 3.13 se presenta una distribución de frecuencias de las calificaciones en un examen final de álgebra. *a*) Encontrar los cuartiles de esta distribución y *b*) interpretar claramente cada uno de ellos.

Tabla 3.13

Calificación	Cantidad de estudiantes
90-100	9
80-89	32
70-79	43
60-69	21
50-59	11
40-49	3
30-39	1
	Total 120

3.108 Encontrar los cuartiles Q_1, Q_2 y Q_3 de las distribuciones: *a*) del problema 3.59 y *b*) del problema 3.60. Interpretar claramente cada uno de ellos.

3.109 Proporcionar seis términos estadísticos diferentes para el punto de equilibrio o valor central en una curva de frecuencias en forma de campana.

3.110 Encontrar: *a*) P_{10}, *b*) P_{90}, *c*) P_{25} y *d*) P_{75} en los datos del problema 3.59. Interpretar claramente cada uno de ellos.

3.111 *a*) ¿Se pueden expresar todos los deciles y cuartiles como percentiles? Explicar.

b) ¿Se pueden expresar los cuantiles como percentiles? Explicar.

3.112 Para los datos del problema 3.107, determinar: *a*) la calificación más baja obtenida por el 25% superior de los alumnos y *b*) la puntuación más alta alcanzada por el 20% inferior de los alumnos. Interpretar las respuestas en términos de percentiles.

3.113 Interpretar gráficamente los resultados del problema 3.107 empleando: *a*) un histograma porcentual, *b*) un polígono de frecuencia porcentual y *c*) una ojiva porcentual.

3.114 Repetir el problema 3.113 para los resultados del problema 3.108.

3.115 *a*) Desarrollar una fórmula similar a la de la ecuación (*8*) de este capítulo que permita calcular cualquier percentil de una distribución de frecuencias.

b) Ilustrar el uso de la fórmula empleándola para obtener los resultados del problema 3.110.

DESVIACIÓN ESTÁNDAR Y OTRAS MEDIDAS DE DISPERSIÓN

4

DISPERSIÓN O VARIACIÓN

El grado de dispersión de los datos numéricos respecto a un valor promedio se llama *dispersión* o *variación* de los datos. Existen varias medidas de dispersión (o variación); las más usadas son el rango, la desviación media, el rango semiintercuartil, el rango percentil 10-90 y la desviación estándar.

RANGO

El *rango* de un conjunto de números es la diferencia entre el número mayor y el número menor del conjunto.

EJEMPLO 1 El rango del conjunto 2, 3, 3, 5, 5, 5, 8, 10, 12 es $12 - 2 = 10$. Algunas veces el rango se da mediante el número menor y el número mayor; así, por ejemplo, en el caso del conjunto anterior, simplemente se indica de 2 a 12 o 2-12.

DESVIACIÓN MEDIA

La *desviación media*, o *desviación promedio*, de un conjunto de N números $X_1, X_2, \ldots, X_N$ se abrevia DM y está definida así:

$$\text{Desviación media (DM)} = \frac{\sum_{j=1}^{N} |X_j - \bar{X}|}{N} = \frac{\sum |X - \bar{X}|}{N} = \overline{|X - \bar{X}|} \qquad (1)$$

donde $\bar{X}$ es la media aritmética de los números y $|X_j - \bar{X}|$ es el valor absoluto de la desviación de X_j respecto de $\bar{X}$. (El valor absoluto de un número es el número sin signo; el valor absoluto de un número se indica por medio de dos barras verticales colocadas a los lados del número, así $|-4| = 4$, $|+3| = 3$, $|6| = 6$ y $|-0.84| = 0.84$|.)

EJEMPLO 2 Encuentre la desviación media del conjunto 2, 3, 6, 8, 11.

$$\text{Media aritmética } (\bar{X}) = \frac{2+3+6+8+11}{5} = 6$$

$$\text{DM} = \frac{|2-6|+|3-6|+|6-6|+|8-6|+|11-6|}{5} = \frac{|-4|+|-3|+|0|+|2|+|5|}{5} = \frac{4+3+0+2+5}{5} = 2.8$$

Si $X_1, X_2, \ldots, X_K$ se presentan con frecuencias $f_1, f_2, \ldots, f_K$, respectivamente, la desviación media puede expresarse como

$$\text{DM} = \frac{\sum_{j=1}^{K} f_j |X_j - \bar{X}|}{N} = \frac{\sum f\,|X - \bar{X}|}{N} = \overline{|X - \bar{X}|} \tag{2}$$

donde $N = \sum_{j=1}^{K} f_j = \sum f$. Esta fórmula es útil para datos agrupados, donde las X_j representan las marcas de clase y las f_j las correspondientes frecuencias de clase.

En ocasiones, la desviación media se define en términos de las desviaciones absolutas respecto de la mediana o de otro promedio, y no respecto de la media. Una propiedad interesante de la suma $\sum_{j=1}^{N} |X_j - a|$ es que es mínima cuando a es la mediana (es decir, la desviación media absoluta con respecto de la mediana es un mínimo).

Obsérvese que sería más apropiado emplear el término *desviación media absoluta* en vez de *desviación media.*

RANGO SEMIINTERCUARTIL

El *rango semiintercuartil*, o *desviación cuartil*, de un conjunto de datos se denota Q y está definido por

$$Q = \frac{Q_3 - Q_1}{2} \tag{3}$$

donde Q_1 y Q_3 son el primero y tercer cuartiles en los datos (ver problemas 4.6 y 4.7). Algunas veces se usa el rango intercuartil $Q_3 - Q_1$; sin embargo, el rango semiintercuartil es más usado como medida de dispersión.

RANGO PERCENTIL 10-90

El *rango percentil 10-90* de un conjunto de datos está definido por

$$\text{Rango percentil 10-90} = P_{90} - P_{10} \tag{4}$$

donde P_{10} y P_{90} son los percentiles 10o. y 90o. en los datos (ver problema 4.8). El rango semipercentil 10-90, $\frac{1}{2}(P_{90} - P_{10})$, también puede usarse, pero no es muy común.

DESVIACIÓN ESTÁNDAR

La *desviación estándar* de un conjunto de N números $X_1, X_2, \ldots, X_N$ se denota como s y está definida por

$$s = \sqrt{\frac{\sum_{j=1}^{N} (X_j - \bar{X})^2}{N}} = \sqrt{\frac{\sum (X - \bar{X})^2}{N}} = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\overline{(X - \bar{X})^2}} \tag{5}$$

donde x representa la desviación de cada uno de los números X_j respecto a la media $\bar{X}$. Por lo tanto, s es la raíz cuadrada de la media (RCM) de las desviaciones respecto de la media, o, como suele llamársele algunas veces, la *desviación raíz-media-cuadrado.*

Si $X_1, X_2, \ldots, X_N$ se presentan con frecuencias $f_1, f_2, \ldots, f_K$, respectivamente, la desviación estándar se puede expresar como

$$s = \sqrt{\frac{\sum_{j=1}^{K} f_j (X_j - \bar{X})^2}{N}} = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}} = \sqrt{\frac{\sum fx^2}{N}} = \sqrt{\overline{(X - \bar{X})^2}} \tag{6}$$

donde $N = \sum_{j=1}^{K} f_j = \sum f$. Esta fórmula es útil para datos agrupados.

Algunas veces la desviación estándar de una muestra de datos se define usando como el denominador, en las ecuaciones (*5*) y (*6*), $(N - 1)$ en lugar de N. Esto se debe a que el valor que así se obtiene es una mejor aproximación a la desviación estándar de la población de la que se ha tomado la muestra. Con valores grandes de N ($N > 30$), prácticamente no hay diferencia entre las dos definiciones. Y cuando se necesita una estimación mejor, ésta siempre se puede obtener multiplicando por $\sqrt{N/(N-1)}$ la desviación estándar obtenida de acuerdo con la primera definición. Por lo tanto, en este libro se emplearán las fórmulas (*5*) y (*6*).

VARIANZA

La *varianza* de un conjunto de datos se define como el cuadrado de la desviación estándar y, por lo tanto, corresponde al valor s^2 en las ecuaciones (*5*) y (*6*).

Cuando es necesario distinguir la desviación estándar de una población de la desviación estándar de una muestra obtenida de esa población, se suele emplear s para la última y σ (letra griega *sigma* minúscula) para la primera. De manera que s^2 y σ^2 representan la *varianza muestral* y *la varianza poblacional*, respectivamente.

MÉTODO ABREVIADO PARA EL CÁLCULO DE LA DESVIACIÓN ESTÁNDAR

Las ecuaciones (*5*) y (*6*) se pueden expresar, respectivamente, mediante las fórmulas siguientes

$$s = \sqrt{\frac{\sum_{j=1}^{N} X_j^2}{N} - \left(\frac{\sum_{j=1}^{N} X_j}{N}\right)^2} = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2} \tag{7}$$

$$s = \sqrt{\frac{\sum_{j=1}^{K} f_j X_j^2}{N} - \left(\frac{\sum_{j=1}^{K} f_j X_j}{N}\right)^2} = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2} \tag{8}$$

donde $\overline{X^2}$ representa la media de los cuadrados de los diversos valores de X, en tanto que $\bar{X}^2$ denota el cuadrado de la media de los diversos valores de X (ver problemas 4.12 a 4.14).

Si las $d_j = X_j - A$ son las desviaciones de X_j respecto a una constante arbitraria A, las fórmulas (*7*) y (*8*) se transforman, respectivamente, en

$$s = \sqrt{\frac{\sum_{j=1}^{N} d_j^2}{N} - \left(\frac{\sum_{j=1}^{N} d_j}{N}\right)^2} = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} = \sqrt{\overline{d^2} - \bar{d}^2} \tag{9}$$

$$s = \sqrt{\frac{\sum_{j=1}^{K} f_j d_j^2}{N} - \left(\frac{\sum_{j=1}^{K} f_j d_j}{N}\right)^2} = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\overline{d^2} - \bar{d}^2} \tag{10}$$

(Ver los problemas 4.15 y 4.17.)

Cuando en una distribución de frecuencia se tienen datos agrupados y los intervalos de clase son de un mismo tamaño c, se tiene $d_j = cu_j$, o $X_j = A + cu_j$ y la fórmula (*10*) se trasforma en

$$s = c\sqrt{\frac{\sum_{j=1}^{K} f_j u_j^2}{N} - \left(\frac{\sum_{j=1}^{K} f_j u_j}{N}\right)^2} = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = c\sqrt{\overline{u^2} - \bar{u}^2} \qquad (11)$$

Esta última fórmula proporciona un método muy sencillo para el cálculo de la desviación estándar y se recomienda su uso para datos agrupados, siempre que los intervalos de clase sean de un mismo tamaño. A este método se le llama *método de compilación* y es exactamente análogo al empleado en el capítulo 3 para calcular la media aritmética de datos agrupados. (Ver problemas 4.16 a 4.19.)

PROPIEDADES DE LA DESVIACIÓN ESTÁNDAR

1. La desviación estándar se puede definir como

$$s = \sqrt{\frac{\sum_{j=1}^{N} (X_j - a)^2}{N}}$$

donde a es un promedio cualquiera además de la media aritmética. De todas las desviaciones estándar, la mínima es aquella en la que $a = \bar{X}$, debido a la propiedad 2 del capítulo 3. Esta propiedad es una razón importante para definir la desviación estándar como se definió antes. En el problema 4.27 se presenta una demostración de esta propiedad.

2. En las distribuciones normales (ver capítulo 7) se encuentra que (como se muestra en la figura 4.1):

   a) 68.27% de los casos está comprendido entre $\bar{X} - s$ y $\bar{X} + s$ (es decir, una desviación estándar a cada lado de la media).
   b) 95.45% de los casos está comprendido entre $\bar{X} - 2s$ y $\bar{X} + 2s$ (es decir, dos desviaciones estándar a cada lado de la media).
   c) 99.73% de los casos está comprendido entre $\bar{X} - 3s$ y $\bar{X} + 3s$ (es decir, tres desviaciones estándar a cada lado de la media).

   En distribuciones moderadamente sesgadas, estos porcentajes se satisfacen de manera aproximada (ver problema 4.24).

3. Supóngase que dos conjuntos que constan de N_1 y N_2 números (o dos distribuciones de frecuencia con frecuencias totales N_1 y N_2) tienen varianzas s_1^2 y s_2^2, respectivamente, y una *misma* media $\bar{X}$. Entonces, la *varianza combinada* o *conjunta* de los dos conjuntos (o de las dos distribuciones de frecuencia) está dada por

$$s^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2} \qquad (12)$$

   Obsérvese que ésta es una media aritmética ponderada de las dos varianzas. Esta fórmula puede generalizarse a tres o más conjuntos.

4. El teorema de Chebyshev establece que para $k > 1$, por lo menos $(1 - (1/k^2)) \times 100\%$ de la distribución de probabilidad de cualquier variable está a no más de k desviaciones estándar de la media. En particular, para $k = 2$, por lo menos $(1 - (1/2^2)) \times 100\%$ o bien 75% de los datos está en el intervalo $(\bar{x} - 2S, \bar{x} + 2S)$; para $k = 3$, por lo menos $(1 - (1/3^2)) \times 100\%$ u 89% de los datos está en el intervalo $(\bar{x} - 3S, \bar{x} + 3S)$, y para $k = 4$, por lo menos $(1 - (1/4^2)) \times 100\%$ o bien 93.75% de los datos está en el intervalo $(\bar{x} - 4S, \bar{x} + 4S)$.

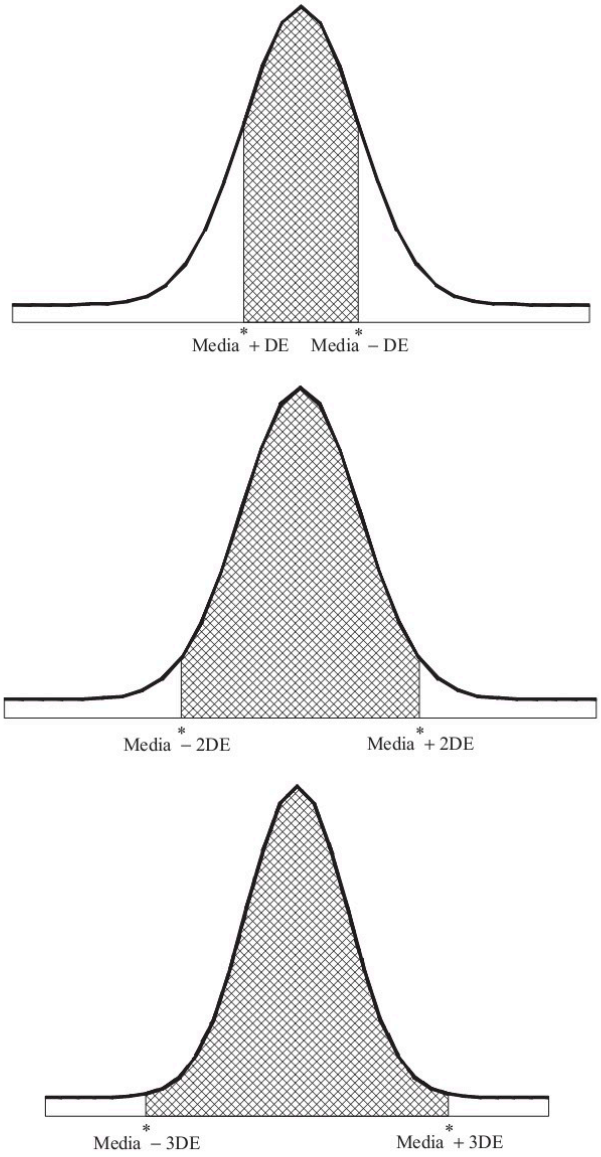


Figura 4-1 Ilustración de la regla empírica.

COMPROBACIÓN DE CHARLIER

La comprobación de Charlier, en el cálculo de la media y de la desviación estándar mediante el método de la compilación, hace uso de las identidades

$$\sum f(u+1) = \sum fu + \sum f = \sum fu + N$$

$$\sum f(u+1)^2 = \sum f(u^2+2u+1) = \sum fu^2 + 2\sum fu + \sum f = \sum fu^2 + 2\sum fu + N$$

(Ver el problema 4.20.)

CORRECCIÓN DE SHEPPARD PARA LA VARIANZA

El cálculo de la desviación estándar tiene cierto error debido a la agrupación de los datos en clases (error de agrupamiento). Para hacer un ajuste respecto al error de agrupamiento, se usa la fórmula

$$\text{Varianza corregida} = \text{Varianza de los datos agrupados} - \frac{c^2}{12} \tag{13}$$

donde c es el tamaño del intervalo de clase. A la corrección $c^2/12$ (que se resta) se le llama *corrección de Sheppard*. Esta corrección se usa para distribuciones de variables continuas, en las que las "colas", en ambas direcciones, se aproximan gradualmente a cero.

Hay discrepancia respecto a *cuándo* y *si* la corrección de Shepppard debe ser aplicada. Desde luego no debe aplicarse antes de que se examine la situación cuidadosamente, ya que se tiende a una *sobrecorrección*, con lo que sólo se sustituye un error por otro. En este libro, a menos que se indique otra cosa, no se usará la corrección de Sheppard.

RELACIONES EMPÍRICAS ENTRE LAS MEDIDAS DE DISPERSIÓN

Para las distribuciones moderadamente sesgadas, se tiene la relación empírica

$$\text{Desviación media} = \tfrac{4}{5}(\text{desviación estándar})$$

$$\text{Rango semiintercuartil} = \tfrac{2}{3}(\text{desviación estándar})$$

Esto es consecuencia de que en una distribución normal se encuentre que la desviación media y el rango semiintercuartil son iguales, respectivamente, a 0.7979 y 0.6745 veces la desviación estándar.

DISPERSIÓN ABSOLUTA Y RELATIVA; COEFICIENTE DE VARIACIÓN

La variación o dispersión real determinada mediante la desviación estándar u otra medida de dispersión se le conoce como *dispersión absoluta*. Sin embargo, una variación o dispersión de 10 pulgadas (in) en una distancia de 1 000 pies (ft) tiene un significado muy diferente a la misma variación de 10 in en una distancia de 20 ft. Este efecto se puede medir mediante la *dispersión relativa*, que se define como sigue:

$$\text{Dispersión relativa} = \frac{\text{dispersión absoluta}}{\text{promedio}} \tag{14}$$

Si la dispersión absoluta es la desviación estándar s y el promedio es la media $\bar{X}$, entonces a la dispersión relativa se le llama *coeficiente de variación* o *coeficiente de dispersión*; este coeficiente se denota por V y está dado por

$$\text{Coeficiente de variación } (V) = \frac{s}{\bar{X}} \tag{15}$$

y por lo general se expresa como porcentaje. También hay otras posibilidades (ver problema 4.30).

Obsérvese que el coeficiente de variación es independiente de las unidades que se empleen. Debido a esto, el coeficiente de variación es útil cuando se trata de comparar distribuciones en las que las unidades son diferentes. Una desventaja del coeficiente de variación es que no es útil cuando el valor de $\bar{X}$ es cercano a cero.

VARIABLE ESTANDARIZADA; PUNTUACIONES ESTÁNDAR

A la variable que mide la desviación respecto a la media en términos de unidades de desviaciones estándar se le llama *variable estandarizada* y es una cantidad adimensional (es decir, es independiente de las unidades empleadas) y está dada por

$$z = \frac{X - \bar{X}}{s} \tag{16}$$

Si las desviaciones respecto a la media se dan en términos de unidades de desviación estándar, se dice que las desviaciones se expresan en *unidades estándar* o en *puntuaciones estándar*. Las unidades estándar son de gran valor para comparar distribuciones (ver problema 4.31).

SOFTWARE Y MEDIDAS DE DISPERSIÓN

El software para estadística proporciona diversas medidas de dispersión. Estas medidas de dispersión suelen proporcionarse en estadística descriptiva. EXCEL permite el cálculo de todas las medidas estudiadas en este libro. Aquí se discuten MINITAB y EXCEL y en los problemas resueltos se muestran los resultados que proporcionan otros paquetes.

EJEMPLO 3

a) EXCEL proporciona cálculos para varias medidas de dipersión, y en el siguiente ejemplo se ilustran algunas de ellas. En una empresa se hace una encuesta; la pregunta es: ¿cuántos e-mails recibe una persona por semana? Las respuestas dadas por los 75 empleados se muestran en las celdas A1:E15 de la hoja de cálculo de EXCEL.

32	113	70	60	84
114	31	58	86	102
113	79	86	24	40
44	42	54	71	25
42	116	68	30	63
121	74	77	77	100
51	31	61	28	26
47	54	74	57	35
77	80	125	105	61
102	45	115	36	52
58	24	24	39	40
95	99	54	35	31
77	29	69	58	32
49	118	44	95	65
71	65	74	122	99

El rango se obtiene mediante =MAX(A1:E15)-MIN(A1:E15) o bien 125 − 24 = 101. La desviación media o desviación promedio se obtiene mediante = DESVPROM(A1:E15) o bien 24.42. El rango semiintercuartil se obtiene mediante la expresión =(PERCENTIL(A1:E15,0.75)-(PERCENTIL(A1:E15,0.25))/2 o bien 22. El rango percentil 10-90 se obtiene mediante PERCENTIL(A1:E15,0.9)-PERCENTIL(A1:E15,0.1) u 82.6.

La desviación estándar y la varianza se obtienen mediante =DESVEST(A1:E15), que es 29.2563 y =VAR(A1:E15), que es 855.932 para muestras, y =DESVESTP(A1:E15) que es 29.0606 y =VARP(A1:E15), que es 844.52 para poblaciones.

b)

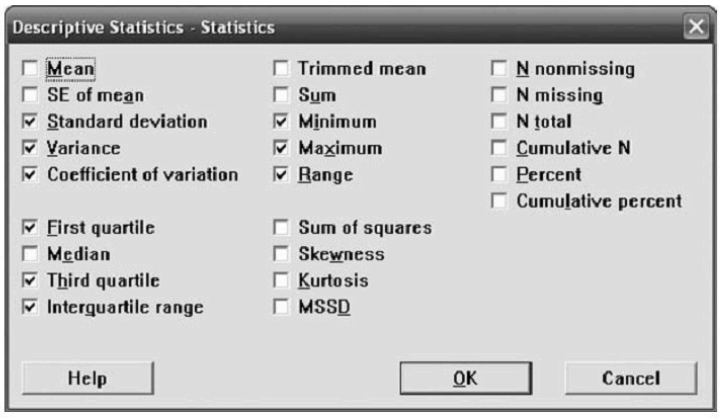


Figura 4-2 Ventana de diálogo de MINITAB.

En la ventana de diálogo de MINITAB, que se presenta en la figura 4-2, se han elegido las medidas de dispersión y de tendencia central. El resultado es el siguiente:

Estadística descriptiva: e-mails

Variable	StDev	Variance	CoefVar	Minimum	Q1	Q3	Maximum	Range	IQR
e-mails	29.26	855.93	44.56	24.00	40.00	86.00	125.00	101.00	46.00

PROBLEMAS RESUELTOS

EL RANGO

4.1 Encontrar el rango de los conjuntos: *a*) 12, 6, 7, 3, 15, 10, 18, 5 y *b*) 9, 3, 8, 8, 9, 8, 9, 18.

SOLUCIÓN

En ambos casos, rango = número mayor − número menor = 18 − 3 = 15. Sin embargo, como se puede ver en las ordenaciones de los conjuntos *a*) y *b*),

a) 3, 5, 6, 7, 10, 12, 15, 18 *b*) 3, 8, 8, 8, 9, 9, 9, 18

en el conjunto *a*) hay mucha más variación que en el conjunto *b*). En efecto, *b*) consta casi únicamente de ochos y nueves.

Dado que el rango no indica diferencia alguna entre estos conjuntos, en este caso no es una buena medida de dispersión. Cuando hay valores extremos, el rango no suele ser una buena medida de la dispersión.

Eliminando los valores extremos, 3 y 18, se logra una mejora. Entonces, el rango del conjunto *a*) es (15 − 5) = 10, en tanto que el rango del conjunto *b*) es (9 − 8) = 1, lo que muestra claramente que en *a*) hay mayor dispersión que en *b*). Sin embargo, el rango no ha sido definido de esta manera. El rango semiintercuartil y el rango percentil 10-90 están concebidos para obtener una medida mejor que el rango mediante la eliminación de los valores extremos.

4.2 Encontrar el rango de las estaturas de los estudiantes de la universidad XYZ dadas en la tabla 2.1.

SOLUCIÓN

Hay dos maneras para definir el rango de datos agrupados.

Primer método

$$\text{Rango} = \text{marca de clase de la clase más alta} - \text{marca de clase de la clase más baja}$$
$$= 73 - 61 = 12 \text{ in}$$

Segundo método

$$\text{Rango} = \text{frontera superior de la clase más alta} - \text{frontera inferior de la clase más baja}$$
$$= 74.5 - 59.5 = 15 \text{ in}$$

Empleando el primer método se tienden a eliminar, en cierta medida, los valores extremos.

LA DESVIACIÓN MEDIA

4.3 Encontrar la desviación media de los conjuntos de números del problema 4.1.

SOLUCIÓN

a) La media aritmética es

$$\bar{X} = \frac{12 + 6 + 7 + 3 + 15 + 10 + 18 + 5}{8} = \frac{76}{8} = 9.5$$

La desviación media es

$$\text{DM} = \frac{\sum |X - \bar{X}|}{N}$$
$$= \frac{|12 - 9.5| + |6 - 9.5| + |7 - 9.5| + |3 - 9.5| + |15 - 9.5| + |10 - 9.5| + |18 - 9.5| + |5 - 9.5|}{8}$$
$$= \frac{2.5 + 3.5 + 2.5 + 6.5 + 5.5 + 0.5 + 8.5 + 4.5}{8} = \frac{34}{8} = 4.25$$

b) $\bar{X} = \dfrac{9 + 3 + 8 + 8 + 9 + 8 + 9 + 18}{8} = \dfrac{72}{8} = 9$

$$\text{DM} = \frac{\sum |X - \bar{X}|}{N}$$
$$= \frac{|9 - 9| + |3 - 9| + |8 - 9| + |8 - 9| + |9 - 9| + |8 - 9| + |9 - 9| + |18 - 9|}{8}$$
$$= \frac{0 + 6 + 1 + 1 + 0 + 1 + 0 + 9}{8} = 2.25$$

La desviación media indica, como debe ser, que en el conjunto *b*) hay menos dispersión que en el conjunto *a*).

4.4 Encontrar la desviación media de las estaturas de 100 estudiantes de la universidad XYZ (ver tabla 3.2, problema 3.20).

SOLUCIÓN

De acuerdo con el problema 3.20, $\bar{X} = 67.45$ in. Para facilitar los cálculos, éstos pueden organizarse como en la tabla 4.1. También se puede idear un método de compilación para el cálculo de la desviación media (ver problema 4.47).

Tabla 4.1

Estaturas (in)	Marcas de clase (X)	$\lvert X - \bar{X}\rvert = \lvert X - 67.45\rvert$	Frecuencia (f)	$f\lvert X - \bar{X}\rvert$
60-62	61	6.45	5	32.25
63-65	64	3.45	18	62.10
66-68	67	0.45	42	18.90
69-71	70	2.55	27	68.85
72-74	73	5.55	8	44.40
			$N = \sum f = 100$	$\sum f\lvert X - \bar{X}\rvert = 226.50$

$$\text{DM} = \frac{\sum f|X - \bar{X}|}{N} = \frac{226.50}{100} = 2.26 \text{ in}$$

4.5 Determinar el porcentaje de las estaturas de los estudiantes del problema 4.4 que cae dentro de los rangos *a*) $\bar{X} \pm \text{DM}$, *b*) $\bar{X} \pm 2$ DM y *c*) $\bar{X} \pm 3$ DM.

SOLUCIÓN

a) El rango de 65.19 a 69.71 in es $\bar{X} \pm \text{DM} = 67.45 \pm 2.26$. Este rango comprende a todos los individuos de la tercera clase $+\frac{1}{3}(65.5 - 65.19)$ de los estudiantes de la segunda clase $+\frac{1}{3}(69.71 - 68.5)$ de los estudiantes de la cuarta clase (ya que el tamaño del intervalo de clase es 3 in, la frontera superior de clase de la segunda clase es 65.5 in y la frontera inferior de clase de la cuarta clase es 68.5 in). La cantidad de estudiantes en el rango $X \pm \text{DM}$ es

$$42 + \frac{0.31}{3}(18) + \frac{1.21}{3}(27) = 42 + 1.86 + 10.89 = 54.75 \qquad \text{o sea} \qquad 55$$

que es 55% del total.

b) El rango de 62.93 a 71.97 in es $\bar{X} \pm 2\text{ DM} = 67.45 \pm 2(2.26) = 67.45 \pm 4.52$. El número de estudiantes en el rango $\bar{X} \pm 2$ DM es

$$18 - \left(\frac{62.93 - 62.5}{3}\right)(18) + 42 + 27 + \left(\frac{71.97 - 71.5}{3}\right)(8) = 85.67 \qquad \text{u} \qquad 86$$

que es 86% del total.

c) El rango de 60.67 a 74.23 in es $\bar{X} \pm 3\text{ DM} = 67.45 \pm 3(2.26) = 67.45 \pm 6.78$. La cantidad de estudiantes en el rango $\bar{X} \pm 3$ DM es

$$5 - \left(\frac{60.67 - 59.5}{3}\right)(5) + 18 + 42 + 27 + \left(\frac{74.5 - 74.23}{3}\right)(8) = 97.33 \qquad \text{o sea} \qquad 97$$

que es 97% del total.

EL RANGO SEMIINTERCUARTIL

4.6 Encontrar el rango semiintercuartil en la distribución de las estaturas de los estudiantes de la universidad XYZ (ver tabla 4.1 del problema 4.4).

SOLUCIÓN

El cuartil inferior y el cuartil superior son $Q_1 = 65.5 + \frac{2}{42}(3) = 65.64$ in y $Q_3 = 68.5 + \frac{10}{27}(3) = 69.61$ in, respectivamente, y el rango semiintercuartil (o desviación cuartil) es $Q = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(69.61 - 65.64) = 1.98$ in. Obsérvese que el 50% de los casos se encuentra entre Q_1 y Q_3 (es decir, la estatura de 50 estudiantes está entre 65.64 y 69.61 in).

Se puede considerar que $\frac{1}{2}(Q_1 + Q_3) = 67.63$ in es una medida de tendencia central (es decir, una altura promedio). Por lo tanto, 50% de las estaturas se encuentra entre 67.63 ± 1.98 in.

4.7 Encontrar el rango semiintercuartil de los salarios de 65 empleados de la empresa P&R (ver la tabla 2.5 del problema 2.3).

SOLUCIÓN

De acuerdo con el problema 3.44, $Q_1 = \$268.25$ y $Q_3 = \$290.75$. Por lo tanto, el rango semiintercuartil es $Q = \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(\$290.75 - \$268.25) = \$11.25$. Como $\frac{1}{2}(Q_1 + Q_3) = \$279.50$, se puede concluir que 50% de los empleados tienen salarios que se encuentran en el rango de $\$279.50 \pm \$11.25$.

EL RANGO PERCENTIL 10-90

4.8 Encontrar el rango percentil 10-90 de las estaturas de los estudiantes de la universidad XYZ (ver tabla 2.1).

SOLUCIÓN

Aquí, $P_{10} = 62.5 + \frac{5}{18}(3) = 63.33$ in y $P_{90} = 68.5 + \frac{25}{27}(3) = 71.27$ in. Por lo tanto, el rango percentil 10-90 es $P_{90} - P_{10} = 71.27 - 63.33 = 7.94$ in. Como $\frac{1}{2}(P_{10} + P_{90}) = 67.30$ in y $\frac{1}{2}(P_{90} - P_{10}) = 3.97$ in, se puede concluir que las estaturas de 80% de los estudiantes se encuentra en el rango de 67.30 ± 3.97 in.

LA DESVIACIÓN ESTÁNDAR

4.9 Encontrar la desviación estándar s de cada uno de los conjuntos de números del problema 4.1.

SOLUCIÓN

a)

$$\bar{X} = \frac{\sum X}{N} = \frac{12+6+7+3+15+10+18+5}{8} = \frac{76}{8} = 9.5$$

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

$$= \sqrt{\frac{(12-9.5)^2+(6-9.5)^2+(7-9.5)^2+(3-9.5)^2+(15-9.5)^2+(10-9.5)^2+(18-9.5)^2+(5-9.5)^2}{8}}$$

$$= \sqrt{23.75} = 4.87$$

b)

$$\bar{X} = \frac{9+3+8+8+9+8+9+18}{8} = \frac{72}{8} = 9$$

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

$$= \sqrt{\frac{(9-9)^2+(3-9)^2+(8-9)^2+(8-9)^2+(9-9)^2+(8-9)^2+(9-9)^2+(18-9)^2}{8}}$$

$$= \sqrt{15} = 3.87$$

Comparando los resultados anteriores con los del problema 4.3 se observa que la desviación estándar sí indica que el conjunto *b*) tiene menos dispersión que el conjunto *a*). Sin embargo, este efecto se enmascara por el hecho de que los valores extremos afectan a la desviación estándar mucho más que a la desviación media. Esto es de esperar, ya que para calcular la desviación estándar las desviaciones se elevan al cuadrado.

4.10 La desviación estándar de los dos conjuntos de datos dados en el problema 4.1 pueden encontrarse con MINITAB. Adelante se presentan los resultados. Comparlos con los obtenidos en el problema 4.9.

```
MTB > print c1
set 1
      12      6      7      3      15      10      18      5
MTB > print c2
set 2
       9      3      8      8       9       8       9     18
MTB > standard deviation c1
```

Columna de desviación estándar

```
   Standard deviation of set 1 = 5.21
MTB > standard deviation c2
```

Columna de desviación estándar

```
   Standard deviation of set 2 = 4.14
```

SOLUCIÓN

MINITAB emplea la fórmula

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}}$$

y por lo tanto, en los problemas 4.9 y 4.10 no se obtiene la misma desviación estándar. Las respuestas del problema 4.10 se pueden obtener de las del problema 4.9 multiplicando éstas por $\sqrt{N/(N-1)}$. Como $N = 8$ para ambos conjuntos, $\sqrt{N/(N-1)} = 1.069045$. Entonces, para el conjunto 1 se tiene (1.069045)(4.87) = 5.21, que es la desviación estándar dada por MINITAB. De igual manera, (1.069045)(3.87) = 4.14, que es la desviación estándar dada por MINITAB para el problema 2.

4.11 Encuentre la desviación estándar de las estaturas de los 100 estudiantes de la universidad XYZ (ver tabla 2.1).

SOLUCIÓN

De acuerdo con los problemas 3.15, 3.20 o bien 3.22, $\bar{X} = 67.45$ in. Los cálculos pueden organizarse como en la tabla 4.2.

$$s = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}} = \sqrt{\frac{852.7500}{100}} = \sqrt{8.5275} = 2.92 \text{ in}$$

Tabla 4.2

Estaturas (in)	Marcas de clase (X)	$X - \bar{X} = X - 67.45$	$(X - \bar{X})^2$	Frecuencias (f)	$f(X - \bar{X})^2$
60-62	61	−6.45	41.6025	5	208.0125
63-65	64	−3.45	11.9025	18	214.2450
66-68	67	−0.45	0.2025	42	8.5050
69-71	70	2.55	6.5025	27	175.5675
72-74	73	5.55	30.8025	8	246.4200
				$N = \sum f = 100$	$\sum f(X - \bar{X})^2 = 852.7500$

CÁLCULO DE LAS DESVIACIONES ESTÁNDAR DE DATOS AGRUPADOS

4.12 *a*) Demostrar que

$$s = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2}$$

b) Usar la fórmula del inciso *a*) para hallar la desviación estándar del conjunto 12, 6, 7, 3, 15, 10, 18, 5.

SOLUCIÓN

a) Por definición

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

Entonces
$$s^2 = \frac{\sum (X - \bar{X})^2}{N} = \frac{\sum (X^2 - 2\bar{X}X + \bar{X}^2)}{N} = \frac{\sum X^2 - 2\bar{X}\sum X + N\bar{X}^2}{N}$$
$$= \frac{\sum X^2}{N} - 2\bar{X}\frac{\sum X}{N} + \bar{X}^2 = \frac{\sum X^2}{N} - 2\bar{X}^2 + \bar{X}^2 = \frac{\sum X^2}{N} - \bar{X}^2$$
$$= \overline{X^2} - \bar{X}^2 = \frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2$$

o bien
$$s = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2}$$

Obsérvese que en las sumatorias anteriores se ha usado la forma abreviada, empleando X en lugar de X_j y $\sum$ en lugar de $\sum_{j=1}^{N}$.

Otro método

$$s^2 = \overline{(X - \bar{X})^2} = \overline{X^2 - 2X\bar{X} + \bar{X}^2} = \overline{X^2} - \overline{2X\bar{X}} + \overline{\bar{X}^2} = \overline{X^2} - 2\bar{X}\bar{X} + \bar{X}^2 = \overline{X^2} - \bar{X}^2$$

b)
$$\overline{X^2} = \frac{\sum X^2}{N} = \frac{(12)^2 + (6)^2 + (7)^2 + (3)^2 + (15)^2 + (10)^2 + (18)^2 + (5)^2}{8} = \frac{912}{8} = 114$$
$$\bar{X} = \frac{\sum X}{N} = \frac{12 + 6 + 7 + 3 + 15 + 10 + 18 + 5}{8} = \frac{76}{8} = 9.5$$

Por lo tanto
$$s = \sqrt{\overline{X^2} - \bar{X}^2} = \sqrt{114 - 90.25} = \sqrt{23.75} = 4.87$$

Compárese este método con el del problema 4.9*a*).

4.13 Modificar la fórmula del problema 4.12*a*) para introducir las frecuencias que corresponden a los diversos valores de X.

SOLUCIÓN

La modificación apropiada es

$$s = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2} = \sqrt{\overline{X^2} - \bar{X}^2}$$

Como en el problema 4.12*a*), a esta fórmula se puede llegar partiendo de

$$s = \sqrt{\frac{\sum f(X - \bar{X})^2}{N}}$$

Entonces
$$s^2 = \frac{\sum f(X-\bar{X})^2}{N} = \frac{\sum f(X^2 - 2\bar{X}X + \bar{X}^2)}{N} = \frac{\sum fX^2 - 2\bar{X}\sum fX + \bar{X}^2 \sum f}{N}$$
$$= \frac{\sum fX^2}{N} - 2\bar{X}\frac{\sum fX}{N} + \bar{X}^2 = \frac{\sum fX^2}{N} - 2\bar{X}^2 + \bar{X}^2 = \frac{\sum fX^2}{N} - \bar{X}^2$$
$$= \frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2$$

o bien
$$s = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2}$$

Obsérvese que la sumatoria anterior se ha usado en forma abreviada, empleando X y f en lugar de X_j y f_j, $\sum$ en lugar de $\sum_{j=1}^{K}$ y $\sum_{j=1}^{K} f_j = N$.

4.14 Empleando la fórmula del problema 4.13, encontrar la desviación estándar de los datos de la tabla 4.2, problema 4.11.

SOLUCIÓN

Los cálculos pueden organizarse como en la tabla 4.3, donde $\bar{X} = (\sum fX)/N = 67.45$ in, según se obtuvo en el problema 3.15. Observar que este método, como el del problema 4.11, conlleva cálculos muy tediosos. En el problema 4.17 se muestra cómo con el método de compilación se simplifican los cálculos enormemente.

Tabla 4.3

Estaturas (in)	Marcas de clase (X)	X^2	Frecuencias (f)	fX^2
60-62	61	3 721	5	18 605
63-65	64	4 096	18	73 728
66-68	67	4 489	42	188 538
69-71	70	4 900	27	132 300
72-74	73	5 329	8	42 632
			$N = \sum f = 100$	$\sum fX^2 = 455\,803$

$$s = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2} = \sqrt{\frac{455\ 803}{100} - (67.45)^2} = \sqrt{8.5275} = 2.92\,\text{in}$$

4.15 Si $d = X - A$ son las desviaciones de X respecto a una constante arbitraria A, probar que
$$s = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

SOLUCIÓN

Como $d = X - A$, $X = A + d$ y $\bar{X} = A + \bar{d}$ (ver problema 3.18), entonces
$$X - \bar{X} = (A + d) - (A + \bar{d}) = d - \bar{d}$$

de manera que
$$s = \sqrt{\frac{\sum f(X-\bar{X})^2}{N}} = \sqrt{\frac{\sum f(d-\bar{d})^2}{N}} = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

de acuerdo con los resultados del problema 4.13 y sustituyendo X y $\bar{X}$ en lugar de d y $\bar{d}$, respectivamente.

Otro método

$$s^2 = \overline{(X - \bar{X})^2} = \overline{(d - \bar{d})^2} = \overline{d^2 - 2\bar{d}d + \bar{d}^2}$$

$$= \overline{d^2} - 2\bar{d}^2 + \bar{d}^2 = \overline{d^2} - \bar{d}^2 = \frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2$$

y la fórmula deseada se obtiene sacando la raíz cuadrada positiva.

4.16 Mostrar que si en una distribución de frecuencia en la que todos los intervalos de clase son del mismo tamaño c, se compila cada marca de clase X con su valor correspondiente u de acuerdo con la relación $X = A + cu$, donde A es una marca de clase dada, entonces la desviación estándar se puede expresar como

$$s = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = c\sqrt{\overline{u^2} - \bar{u}^2}$$

SOLUCIÓN

Esto se deduce inmediatamente del problema 4.15, ya que $d = X - A = cu$. Por lo tanto, como c es una constante,

$$s = \sqrt{\frac{\sum f(cu)^2}{N} - \left(\frac{\sum f(cu)}{N}\right)^2} = \sqrt{c^2\frac{\sum fu^2}{N} - c^2\left(\frac{\sum fu}{N}\right)^2} = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2}$$

Otro método

Esta fórmula se puede probar también directamente sin usar el problema 4.15. Dado que $X = A + cu$, $\bar{X} = A + c\bar{u}$ y $X - \bar{X} = c(u - \bar{u})$, entonces

$$s^2 = \overline{(X - \bar{X})^2} = \overline{c^2(u - \bar{u})^2} = c^2\overline{(u^2 - 2\bar{u}u + \bar{u}^2)} = c^2(\overline{u^2} - 2\bar{u}^2 + \bar{u}^2) = c^2(\overline{u^2} - \bar{u}^2)$$

y

$$s = c\sqrt{\overline{u^2} - \bar{u}^2} = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2}$$

4.17 Encontrar la desviación estándar de las estaturas de los estudiantes de la universidad XYZ (ver la tabla 2.1) empleando: *a*) la fórmula obtenida en el problema 4.15 y *b*) el método de codificación del problema 4.16.

SOLUCIÓN

En las tablas 4.4 y 4.5 arbitrariamente se ha elegido A igual a la marca de clase 67. Obsérvese que en la tabla 4.4 las desviaciones $d = X - A$ son múltiplos del tamaño del intervalo de clase $c = 3$. En la tabla 4.5 se ha eliminado este factor. Esto da como resultado que en la tabla 4.5 los cálculos se simplifican enormemente (en comparación con los de los problemas 4.11 y 4.14). Por esto se recomienda emplear el método de compilación siempre que sea posible.

a) Ver la tabla 4.4.

Tabla 4.4

Marcas de clase (X)	$d = X - A$	Frecuencias (f)	fd	fd^2
61	−6	5	−30	180
64	−3	18	−54	162
$A \rightarrow 67$	0	42	0	0
70	3	27	81	243
73	6	8	48	288
		$N = \sum f = 100$	$\sum fd = 45$	$\sum fX^2 = 873$

$$s = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{873}{100} - \left(\frac{45}{100}\right)^2} = \sqrt{8.5275} = 2.92 \text{ in}$$

b) Ver la tabla 4.5

Tabla 4.5

Marcas de clase (X)	$u = \dfrac{X - A}{c}$	Frecuencias (f)	fu	fu^2
61	−2	5	−10	20
64	−2	18	−18	18
$A \rightarrow 67$	0	42	0	0
70	1	27	27	27
73	2	8	18	32
		$N = \sum f = 100$	$\sum fu = 15$	$\sum fu^2 = 97$

$$s = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = 3\sqrt{\frac{97}{100} - \left(\frac{15}{100}\right)^2} = 3\sqrt{0.9475} = 2.92 \text{ in}$$

4.18 Empleando el método de compilación, encontrar: *a*) la media y *b*) la desviación estándar de la distribución de los salarios de los 65 empleados de la empresa P&R (ver la tabla 2.5 del problema 2.3).

SOLUCIÓN

Los cálculos se pueden organizar como en la tabla 4.6.

a) $$\bar{X} = A + c\bar{u} = A + c\,\frac{\sum fu}{N} = \$275.00 + (\$10.00)\left(\frac{31}{65}\right) = \$279.77$$

b) $$s = c\sqrt{\overline{u^2} - \bar{u}^2} = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = (\$10.00)\sqrt{\frac{173}{65} - \left(\frac{31}{65}\right)^2} = (\$10.00)\sqrt{2.4341} = \$15.60$$

Tabla 4.6

	X	u	f	fu	fu^2
	\$255.00	−2	8	−16	32
	265.00	−1	10	−10	10
$A \longrightarrow$	275.00	0	16	0	0
	285.00	1	14	14	14
	295.00	2	10	20	40
	305.00	3	5	15	45
	315.00	4	2	8	32
			$N = \sum f = 65$	$\sum fu = 31$	$\sum fu^2 = 173$

4.19 La tabla 4.7 muestra el CI de 480 niños de primaria. Empleando el método de compilación, encontrar: *a*) la media y *b*) la desviación estándar.

Tabla 4.7

Marca des clase (X)	70	74	78	82	86	90	94	98	102	106	110	114	118	122	126
Frecuencias (f)	4	9	16	28	45	66	85	72	54	38	27	18	11	5	2

SOLUCIÓN

El cociente intelectual es

$$\text{CI} = \frac{\text{edad mental}}{\text{edad cronológica}}$$

expresado como porcentaje. Por ejemplo, un niño de 8 años que (de acuerdo con ciertos procedimientos educativos) tiene una mentalidad de un niño de 10 años, tendrá un CI de $10/8 = 1.25 = 125\%$, o simplemente 125, el signo % se sobreentiende.

Para hallar la media y la desviación estándar de los cocientes intelectuales de la tabla 4.7, se pueden organizar los cálculos como en la tabla 4.8.

a) $$\bar{X} = A + c\bar{u} = A + c\,\frac{\sum fu}{N} = 94 + 4\left(\frac{236}{480}\right) = 95.97$$

b) $$s = c\sqrt{\bar{u^2} - \bar{u}^2} = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = 4\sqrt{\frac{3\,404}{480} - \left(\frac{236}{480}\right)^2} = 4\sqrt{6.8499} = 10.47$$

COMPROBACIÓN DE CHARLIER

4.20 Emplear la comprobación de Charlier para verificar los cálculos de: *a*) la media y *b*) la desviación estándar realizados en el problema 4.19.

SOLUCIÓN

Para hacer la comprobación deseada, a las columnas de la tabla 4.8 se agregan las columnas de la tabla 4.9 (con excepción de la columna 2, que por comodidad se repite en la tabla 4.9).

a) De acuerdo con la tabla 4.9, $\sum f(u+1) = 716$; de acuerdo con la tabla 4.8, $\sum fu + N = 236 + 480 = 716$. Con esto se tiene la comprobación de la media.

Tabla 4.8

	X	u	f	fu	fu^2
	70	−6	4	−24	144
	74	−5	9	−45	225
	78	−4	16	−64	256
	82	−3	28	−84	252
	86	−2	45	−90	180
	90	−1	66	−66	66
$A \longrightarrow$	94	0	85	0	0
	98	1	72	72	72
	102	2	54	108	216
	106	3	38	114	342
	110	4	27	108	432
	114	5	18	90	450
	118	6	11	66	396
	122	7	5	35	245
	126	8	2	16	128
			$N = \sum f = 480$	$\sum fu = 236$	$\sum fu^2 = 3\,404$

Tabla 4.9

$u+1$	f	$f(u+1)$	$f(u+1)^2$
−5	4	−20	100
−4	9	−36	144
−3	16	−48	144
−2	28	−56	112
−1	45	−45	45
0	66	0	0
1	85	85	85
2	72	144	288
3	54	162	486
4	38	152	608
5	27	135	675
6	18	108	648
7	11	77	539
8	5	40	320
9	2	18	162
	$N=\sum f=480$	$\sum f(u+1)=716$	$\sum f(u+1)^2=4\,356$

b) De acuerdo con la tabla 4.9, $\sum f(u+1)^2 = 4\,356$; de acuerdo con la tabla 4.8, $\sum f^2 + 2\sum fu + N = 3\,404 + 2(236) + 480 = 4\,356$, con lo que se tiene la comprobación de la desviación estándar.

CORRECCIÓN DE SHEPPARD PARA LA VARIANZA

4.21 Emplee la corrección de Sheppard para determinar la desviación estándar de los datos en: *a*) el problema 4.17, *b*) el problema 4.18 y *c*) el problema 4.19.

SOLUCIÓN

a) $s^2 = 8.5275$ y $c = 3$. Varianza corregida $= s^2 - c^2/12 = 8.5275 - 3^2/12 = 7.7775$. Desviación estándar corregida $= \sqrt{\text{varianza corregida}} = \sqrt{7.7775} = 2.79$ in.

b) $s^2 = 243.41$ y $c = 10$. Varianza corregida $= s^2 - c^2/12 = 243.41 - 10^2/12 = 235.08$. Desviación estándar corregida $= \sqrt{235.08} = \$15.33$.

c) $s^2 = 109.60$ y $c = 4$. Varianza corregida $= s^2 - c^2/12 = 109.60 - 4^2/12 = 108.27$. Desviación estándar corregida $= \sqrt{108.27} = 10.41$.

4.22 Dada la segunda distribución de frecuencia del problema 2.8, encontrar: *a*) la media, *b*) la desviación estándar, *c*) la desviación estándar usando la corrección de Sheppard y *d*) la verdadera desviación estándar a partir de los datos no agrupados.

SOLUCIÓN

Los cálculos se pueden organizar como en la tabla 4.10.

a)
$$\bar{X} = A + c\bar{u} = A + c\frac{\sum fu}{N} = 149 + 9\left(\frac{-9}{40}\right) = 147.0\,\text{lb}$$

b)
$$s = c\sqrt{\bar{u^2} - \bar{u}^2} = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = 9\sqrt{\frac{95}{40} - \left(\frac{-9}{40}\right)^2} = 9\sqrt{2.324375} = 13.7\,\text{lb}$$

c) Varianza corregida $= s^2 - c^2/12 = 188.27 - 9^2/12 = 181.52$. Desviación estándar corregida $= 13.5$ lb.

Tabla 4.10

	X	u	f	fu	fu^2
	122	−3	3	−9	27
	131	−2	5	−10	20
	140	−1	9	−9	9
$A \longrightarrow$	149	0	12	0	0
	158	1	5	5	5
	167	2	4	8	16
	176	3	2	6	18
			$N = \sum f = 40$	$\sum fu = -9$	$\sum fu^2 = 95$

d) Para calcular la desviación estándar a partir de los verdaderos pesos de los estudiantes, dados en el problema, conviene primero restarle a cada peso un número adecuado, por ejemplo, $A = 150$ lb, y después usar el método del problema 4.15. Las desviaciones $d = X - A = X - 150$ se dan en la tabla siguiente:

−12	14	0	−18	−6	−25	−1	7
−4	8	−10	−3	−14	−2	2	−6
18	−24	−12	26	13	−31	4	15
−4	23	−8	−3	−15	3	−10	−15
11	−5	−15	−8	0	6	−5	−22

a partir de las cuales se encuentra que $\sum d = -128$ y $\sum d^2 = 7\,052$. Entonces

$$s = \sqrt{\overline{d^2} - \bar{d}^2} = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2} = \sqrt{\frac{7\,052}{40} - \left(\frac{-128}{40}\right)^2} = \sqrt{166.06} = 12.9 \text{ lb}$$

Por lo tanto, con la corrección de Sheppard, en este caso, se obtiene cierta mejora.

RELACIONES EMPÍRICAS ENTRE LAS MEDIDAS DE DISPERSIÓN

4.23 Dada la distribución de las estaturas de los estudiantes de la universidad XYZ, comentar la validez de las fórmulas empíricas: *a*) desviación media $= \frac{4}{5}$(desviación estándar) y *b*) rango semiintercuartil $= \frac{2}{3}$(desviación estándar).

SOLUCIÓN

a) De acuerdo con los problemas 4.4 y 4.11, desviación media ÷ desviación estándar $= 2.26/2.92 = 0.77$, que es aproximadamente $\frac{4}{5}$.

b) De acuerdo con los problemas 4.6 y 4.11, rango semiintercuartil ÷ desviación estándar $= 1.98/2.92 = 0.68$, que es aproximadamente $\frac{2}{3}$.

Por lo tanto, en este caso las fórmulas empíricas son válidas.

Obsérvese que no se usó la desviación estándar con corrección de Sheppard para agrupamiento, ya que no se hicieron las correcciones correspondientes a la desviación media ni al rango semiintercuartílico.

PROPIEDADES DE LA DESVIACIÓN ESTÁNDAR

4.24 En el problema 4.19 determinar el porcentaje de estudiantes cuyo CI cae dentro de los rangos: *a*) $\bar{X} \pm s$, *b*) $\bar{X} \pm 2s$ y *c*) $\bar{X} \pm 3s$.

SOLUCIÓN

a) El rango para los CI de 85.5 a 106.4 es $\bar{X} \pm s = 95.97 \pm 10.47$. La cantidad de CI en el rango $\bar{X} \pm s$ es

$$\left(\frac{88 - 85.5}{4}\right)(45) + 66 + 85 + 72 + 54 + \left(\frac{106.4 - 104}{4}\right)(38) = 339$$

El porcentaje de CI en el rango $\bar{X} \pm s$ es $339/480 = 70.6\%$.

b) El rango de los CI de 75.0 a 116.9 es $\bar{X} \pm 2s = 95.97 \pm 2(10.47)$. La cantidad de CI en el rango $\bar{X} \pm 2s$ es

$$\left(\frac{76 - 75.0}{4}\right)(9) + 16 + 28 + 45 + 66 + 85 + 72 + 54 + 38 + 27 + 18 + \left(\frac{116.9 - 116}{4}\right)(11) = 451$$

El porcentaje de CI en el rango $\bar{X} \pm 2s$ es $451/480 = 94.0\%$.

c) El rango de los CI de 64.6 a 127.4 es $\bar{X} \pm 3s = 95.97 \pm 3(10.47)$. La cantidad de CI en el rango $\bar{X} \pm 3s$ es

$$480 - \left(\frac{128 - 127.4}{4}\right)(2) = 479.7 \qquad \text{o} \qquad 480$$

El porcentaje de CI en el rango $\bar{X} \pm 3s$ es $479.7/480 = 100\%$.

Los porcentajes de los incisos *a*), *b*) y *c*) coinciden con los esperados en una distribución normal: 68.27%, 95.45% y 99.73%, respectivamente.

Obsérvese que no se ha usado la corrección de Sheppard para la desviación estándar. Si se usa esta corrección, los resultados, en este caso, coinciden estrechamente con los anteriores. Obsérvese que los resultados anteriores también pueden obtenerse usando la tabla 4.11 del problema 4.32.

4.25 Dados los conjuntos 2, 5, 8, 11, 14 y 2, 8, 14, encontrar: *a*) la media de cada conjunto, *b*) la varianza de cada conjunto, *c*) la media de los conjuntos combinados (o conjuntados) y *d*) la varianza de los conjuntos combinados.

SOLUCIÓN

a) Media del primer conjunto $= \frac{1}{5}(2 + 5 + 8 + 11 + 14) = 8$. Media del segundo conjunto $= \frac{1}{3}(2 + 8 + 14) = 8$.

b) Varianza del primer conjunto $= s_1^2 = \frac{1}{5}[(2 - 8)^2 + (5 - 8)^2 + (8 - 8)^2 + (11 - 8)^2 + (14 - 8)^2] = 18$. Varianza del segundo conjunto $= s_2^2 = \frac{1}{3}[(2 - 8)^2 + (8 - 8)^2 + (14 - 8)^2] = 24$.

c) La media de los conjuntos combinados es

$$\frac{2 + 5 + 8 + 11 + 14 + 2 + 8 + 14}{5 + 3} = 8$$

d) La varianza de los conjuntos combinados es

$$s^2 = \frac{(2 - 8)^2 + (5 - 8)^2 + (8 - 8)^2 + (11 - 8)^2 + (14 - 8)^2 + (2 - 8)^2 + (8 - 8)^2 + (14 - 8)^2}{5 + 3} = 20.25$$

Otro método (mediante fórmula)

$$s^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2} = \frac{(5)(18) + (3)(24)}{5 + 3} = 20.25$$

4.26 Resolver el problema 4.25 con los conjuntos 2, 5, 8, 11, 14 y 10, 16, 22.

SOLUCIÓN

Aquí las medias de los dos conjuntos son 8 y 16, respectivamente, en tanto que las varianzas son las *mismas* que las varianzas en el problema anterior, a saber: $s_1^2 = 18$ y $s_2^2 = 24$.

$$\text{Media de los conjuntos combinados} = \frac{2+5+8+11+14+10+16+22}{5+3} = 11$$

$$s^2 = \frac{(2-11)^2+(5-11)^2+(8-11)^2+(11-11)^2+(14-11)^2+(10-11)^2+(16-11)^2+(22-11)^2}{5+3}$$

$$= 35.25$$

Obsérvese que la fórmula

$$s^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2}$$

con la que se obtiene el valor 20.25, *no* es aplicable en este caso, ya que las medias de los dos conjuntos *no* son iguales.

4.27 *a*) Probar que $w^2 + pw + q$, donde p y q son constantes dadas, es mínimo si y sólo si $w = -\frac{1}{2}p$.

b) Empleando el inciso *a*), probar que

$$\frac{\sum_{j=1}^{N}(X_j - a)^2}{N} \qquad \text{o brevemente} \qquad \frac{\sum(X-a)^2}{N}$$

es mínimo si y sólo si $a = \bar{X}$.

SOLUCIÓN

a) Se tiene $w^2 + pw + q = (w + \frac{1}{2}p)^2 + q - \frac{1}{4}p^2$. Como $(q - \frac{1}{4}p^2)$ es constante, esta expresión tiene su mínimo valor si y sólo si $w + \frac{1}{2}p = 0$ (es decir, $w = -\frac{1}{2}p$).

b)
$$\frac{\sum(X-a)^2}{N} = \frac{\sum(X^2 - 2aX + a^2)}{N} = \frac{\sum X^2 - 2a\sum X + Na^2}{N} = a^2 - 2a\frac{\sum X}{N} + \frac{\sum X^2}{N}$$

Comparando esta última expresión con $(w^2 + pw + q)$, se tiene

$$w = a \qquad p = -2\frac{\sum X}{N} \qquad q = \frac{\sum X^2}{N}$$

Por lo tanto, la expresión tiene un mínimo en $a = -\frac{1}{2}p = (\sum X)/N = \bar{X}$, empleando el resultado del inciso *a*).

DISPERSIÓN ABSOLUTA Y RELATIVA; COEFICIENTE DE VARIACIÓN

4.28 Un fabricante de cinescopios produce dos tipos de cinescopios, *A* y *B*. La vida media de los cinescopios es, respectivamente, $\bar{X}_A = 1\,495$ horas y $\bar{X}_B = 1\,875$ horas, y las desviaciones estándar son $s_A = 280$ horas y $s_B = 310$ horas. ¿Cuál de los cinescopios tiene: *a*) la mayor dispersión absoluta y *b*) la mayor dispersión relativa?

SOLUCIÓN

a) La dispersión absoluta de A es $s_A = 280$ horas y la de B es $s_B = 310$ horas. Por lo tanto, en los cinescopios B hay mayor dispersión absoluta.

b) Los coeficientes de variación son

$$A = \frac{s_A}{\bar{X}_A} = \frac{280}{1\,495} = 18.7\% \qquad B = \frac{s_B}{\bar{X}_B} = \frac{310}{1\,875} = 16.5\%$$

Por lo tanto, los cinescopios A tienen mayor variación relativa o dispersión.

4.29 Encontrar el coeficiente de variación, V, de los datos: *a*) del problema 4.14 y *b*) del problema 4.18, empleando la desviación estándar corregida y la desviación estándar no corregida.

SOLUCIÓN

a)
$$V(\text{no corregida}) = \frac{s(\text{no corregida})}{\bar{X}} = \frac{2.92}{67.45} = 0.0433 = 4.3\%$$

$$V(\text{corregida}) = \frac{s(\text{corregida})}{\bar{X}} = \frac{2.79}{67.45} = 0.0413 = 4.1\% \qquad \text{de acuerdo con el problema 4.21}a)$$

b)
$$V(\text{no corregida}) = \frac{s(\text{no corregida})}{\bar{X}} = \frac{15.60}{79.77} = 0.196 = 19.6\%$$

$$V(\text{corregida}) = \frac{s(\text{corregida})}{\bar{X}} = \frac{15.33}{79.77} = 0.192 = 19.2\% \qquad \text{de acuerdo con el problema 4.21}b)$$

4.30 *a*) Definir una medida de dispersión relativa que pueda emplearse para un conjunto de datos en el que se conocen los cuartiles.

b) Ilustrar el cálculo de la medida definida en el inciso *a*) aplicándolo a los datos del problema 4.6.

SOLUCIÓN

a) Si para un conjunto de datos, se dan los cuartiles Q_1 y Q_3, entonces $\frac{1}{2}(Q_1 + Q_3)$ es una medida de tendencia central de los datos o promedio, en tanto que $Q = \frac{1}{2}(Q_3 - Q_1)$, el rango semiintercuartil, es una medida de dispersión de los datos. De manera que una medida de dispersión relativa se puede definir de la siguiente manera.

$$V_Q = \frac{\frac{1}{2}(Q_3 - Q_1)}{\frac{1}{2}(Q_1 + Q_3)} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

a la que se le llama *coeficiente de variación cuartil* o *coeficiente cuartil de dispersión relativa.*

b)
$$V_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{69.61 - 65.64}{69.61 + 65.64} = \frac{3.97}{135.25} = 0.0293 = 2.9\%$$

VARIABLES ESTANDARIZADAS; PUNTUACIONES ESTÁNDAR

4.31 En el examen final de matemáticas en el que la media es 76 y la desviación estándar es 10, un alumno obtiene una calificación de 84. En el examen final de física, en el que la media es 82 y la desviación estándar es 16, el mismo alumno obtiene como puntuación 90. ¿En qué materia tiene una posición relativa más alta?

SOLUCIÓN

La variable estandarizada $z = (X - \bar{X})/s$ mide la desviación de X respecto a la media $\bar{X}$ en término de desviaciones estándar s. En matemáticas, $z = (84 - 76)/10 = 0.8$, y en física $z = (90 - 82)/16 = 0.5$. Por lo tanto, la calificación de este estudiante en matemáticas se encuentra a 0.8 de una desviación estándar sobre la media, en cambio la puntuación en física se encuentra a sólo 0.5 de una desviación estándar sobre la media. Por lo tanto, en matemáticas obtuvo una posición relativa más alta.

La variable $z = (X - \bar{X})/s$ suele emplearse para las calificaciones de los exámenes de conocimientos, en donde se denomina *calificación estándar.*

SOFTWARE Y MEDIDAS DE DISPERSIÓN

4.32 El análisis hecho con STATISTIX de los datos del ejemplo 3 de este capítulo da los resultados siguientes:

```
Statistix 8.0
Descriptive Statistics

Variable         SD       Variance        C.V.        MAD
e - mails     29.256         855.93      44.562     21.000
```

El valor MAD es la *desviación mediana absoluta*. Se trata del valor mediano de las diferencias absolutas entre cada uno de los valores y la mediana muestral. Verificar que el valor MAD de estos datos es 21.

SOLUCIÓN

Los datos ordenados de menor a mayor son:

24	24	24	25	26	28	29	30	31	31	31	32	32
35	35	36	39	40	40	42	42	44	44	45	47	49
51	52	54	54	54	57	58	58	58	60	61	61	63
65	65	68	69	70	71	71	74	74	74	77	77	77
77	79	80	84	86	86	95	95	99	99	100	102	102
105	113	113	114	115	116	118	121	122	125			

La mediana de los datos originales es 61.

Si a cada dato se le resta 61, se obtiene:

−37	−37	−37	−36	−35	−33	−32	−31	−30	−30	−30	−29	−29
−26	−26	−25	−22	−21	−21	−19	−19	−17	−17	−16	−14	−12
−10	−9	−7	−7	−7	−4	−3	−3	−3	−1	0	0	2
4	4	7	8	9	10	10	13	13	13	16	16	16
16	18	19	23	25	25	34	34	38	38	39	41	41
44	52	52	53	54	55	57	60	61	64			

Ahora se toma el valor absoluto de estos datos:

37	37	37	36	35	33	32	31	30	30	30	29	29	26	26
25	22	21	21	19	19	17	17	16	14	12	10	9	7	7
7	4	3	3	3	1	0	0	2	4	4	7	8	9	10
10	13	13	13	16	16	16	16	18	19	23	25	25	34	34
38	38	39	41	41	44	52	52	53	54	55	57	60	61	64

La mediana de este último conjunto es 21. Por lo tanto, MAD = 21.

PROBLEMAS SUPLEMENTARIOS

RANGO

4.33 Encontrar el rango de los conjuntos: *a*) 5, 3, 8, 4, 7, 6, 12, 4, 3 y *b*) 8.772, 6.453, 10.624, 8.628, 9.434, 6.351.

4.34 Encontrar el rango de las cargas máximas dadas en la tabla 3.8 del problema 3.59.

4.35 Encontrar el rango de los diámetros de remaches dados en la tabla 3.10 del problema 3.61.

4.36 En 50 medidas, la mayor es de 8.34 kilogramos (kg). Si el rango es 0.46 kg, encontrar la medida menor.

4.37 En la tabla siguiente se dan las semanas que necesitaron 25 trabajadores, que perdieron su trabajo por reducción de personal en sus empresas, para encontrar un nuevo empleo. Encontrar el rango de estos datos.

13	13	17	7	22
22	26	17	13	14
16	7	6	18	20
10	17	11	10	15
16	8	16	21	11

DESVIACIÓN MEDIA

4.38 Encontrar el valor absoluto de: *a*) -18.2, *b*) $+3.58$, *c*) 6.21, *d*) 0, *e*) $-\sqrt{2}$ y *f*) $4.00 - 2.36 - 3.52$.

4.39 Encontrar la desviación media de los conjuntos: *a*) 3, 7, 9, 5 y *b*) 2.4, 1.6, 3.8, 4.1, 3.4.

4.40 Encontrar a desviación media de los conjuntos de números del problema 4.33.

4.41 Encontrar la desviación media de las cargas máximas dadas en la tabla 3.8 del problema 3.59.

4.42 *a*) Encontrar la desviación media (DM) de los diámetros de los remaches de la tabla 3.10 del problema 3.61.

b) ¿Qué porcentaje de los diámetros de los remaches está entre ($\bar{X} \pm$ DM), ($\bar{X} \pm 2$ DM) y ($\bar{X} = 3$ DM)?

4.43 En el conjunto 8, 10, 9, 12, 4, 8, 2, encontrar la desviación media: *a*) respecto a la media y *b*) respecto a la mediana. Verificar que la desviación media respecto a la mediana no es mayor que la desviación media respecto a la media.

4.44 En la distribución dada en la tabla 3.9 del problema 3.60, encontrar la desviación media: *a*) respecto a la media y *b*) respecto a la mediana. Emplear los resultados de los problemas 3.60 y 3.70.

4.45 En la distribución dada en la tabla 3.11 del problema 3.62, encontrar la desviación media: *a*) respecto a la media y *b*) respecto a la mediana. Emplear los resultados de los problemas 3.62 y 3.72.

4.46 Encontrar la desviación media de los datos dados en el problema 4.37.

4.47 Deducir fórmulas de compilación para el cálculo de la desviación media: *a*) respecto a la media y *b*) respecto a la mediana a partir de una distribución de frecuencias. Emplear estas fórmulas para verificar los resultados obtenidos en los problemas 4.44 y 4.45.

EL RANGO SEMIINTERCUARTIL

4.48 Encontrar el rango semiintercuartil en las distribuciones: *a*) del problema 3.59, *b*) del problema 3.60 y *c*) del problema 3.107. En cada caso interpretar los resultados claramente.

4.49 Encontrar el rango semiintercuartil de los datos dados en el problema 4.37.

4.50 Probar que en cualquier distribución de frecuencias, el porcentaje de casos que cae en el intervalo $\frac{1}{2}(Q_1 + Q_3) \pm \frac{1}{2}(Q_3 - Q_1)$ es el 50%. ¿Ocurre lo mismo en el intervalo $Q_2 \pm \frac{1}{2}(Q_3 - Q_1)$? Explicar la respuesta.

4.51 *a*) ¿Cómo se graficaría el rango semiintercuartil correspondiente a una distribución de frecuencias dada?

b) ¿Qué relación hay entre el rango semiintercuartil y la ojiva de una distribución?

EL RANGO PERCENTIL 10-90

4.52 Encontrar el rango percentil 10-90 en las distribuciones: *a*) del problema 3.59 y *b*) del problema 3.107. En cada caso interpretar los resultados claramente.

4.53 El décimo percentil de los precios de venta de las casas en determinada ciudad es \$35 500 y el nonagésimo percentil de los precios de venta de las casas en la misma ciudad es \$225 000. Encontrar el rango percentil 10-90 y dar un rango en el que caiga el 80% de los precios de venta.

4.54 ¿Qué ventajas o desventajas tiene un rango percentil 20-80 en comparación con un rango percentil 10-90?

4.55 Contestar el problema 4.51 en relación: *a*) con el rango percentil 10-90, *b*) con el rango percentil 20-80 y *c*) el rango percentil 25-75. ¿Cuál es la relación entre *c*) y el rango semiintercuartil?

LA DESVIACIÓN ESTÁNDAR

4.56 Encontrar la desviación estándar de los conjuntos: *a*) 3, 6, 2, 1, 7, 5; *b*) 3.2, 4.6, 2.8, 5.2, 4.4, y *c*) 0, 0, 0, 0, 0, 1, 1, 1.

4.57 *a*) Sumando 5 a cada uno de los números del conjunto 3, 6, 2, 1, 7, 5 se obtiene el conjunto 8, 11, 7, 6, 12, 10. Mostrar que los dos conjuntos tienen la misma desviación estándar pero diferentes medias. ¿Qué relación hay entre las medias?

b) Si cada uno de los números del conjunto 3, 6, 2, 1, 7 y 5 se multiplica por 2 y después se le suma 5, se obtiene el conjunto 11, 17, 9, 7, 19, 15. ¿Qué relación existe entre las medias y las desviaciones estándar de estos dos conjuntos?

c) ¿Qué propiedades de la media y de la desviación estándar se ilustran mediante los conjuntos de números particulares de los incisos *a*) y *b*)?

4.58 Encontrar la desviación estándar del conjunto de números de la progresión aritmética 4, 10, 16, 22, . . . , 154.

4.59 Encontrar la desviación estándar en las distribuciones: *a*) del problema 3.59, *b*) del problema 3.60 y *c*) del problema 3.107,

4.60 Ilustrar el uso de la comprobación de Charlier en cada inciso del problema 4.59.

4.61 Encontrar: *a*) la media y *b*) la desviación estándar en la distribución del problema 2.17 y explicar el significado de los resultados obtenidos.

4.62 Cuando los datos tienen una distribución en forma de campana, la desviación estándar se puede obtener de manera aproximada dividiendo el rango entre 4. Con los datos dados en el problema 4.37, calcular la desviación estándar y compararla con el rango dividido entre 4.

4.63 *a*) Encontrar la desviación estándar s de los diámetros de los remaches dados en la tabla 3.10 del problema 3.61.
b) ¿Qué porcentaje de los diámetros de los remaches se encuentra entre $\bar{X} \pm s$, $\bar{X} \pm 2s$ y $\bar{X} \pm 3s$?
c) Comparar los porcentajes del inciso *b*) con los que teóricamente se esperan en una distribución normal y explicar cualquier diferencia observada.

4.64 Aplicar la corrección de Sheppard a las desviaciones estándar del problema 4.59. En cada caso, comentar si la aplicación de la corrección de Sheppard está o no justificada.

4.65 ¿Qué modificaciones ocurren en el problema 4.63 cuando se aplica la corrección de Sheppard?

4.66 *a*) Encontrar la media y la desviación estándar de los datos del problema 2.8.
b) Construir una distribución de frecuencia para los datos y encontrar la desviación estándar.
c) Comparar los resultados del inciso *b*) con los del inciso *a*). Determinar si la aplicación de la corrección de Sheppard produce mejores resultados.

4.67 Repetir el problema 4.66 con los datos del problema 2.27.

4.68 *a*) De un total de N números, la fracción p es de unos y la fracción $q = 1 - p$ es de ceros. Probar que la desviación estándar de este conjunto de números es $\sqrt{pq}$.
b) Aplicar el resultado del inciso *a*) al problema 4.56*c*).

4.69 *a*) Probar que la varianza del conjunto de números a, $a + d$, $a + 2d, \ldots, a + (n-1)d$ (es decir, de una progresión aritmética en la que el primer término es a y la diferencia común es d) es $\frac{1}{12}(n^2 - 1)d^2$.
b) Emplear el inciso *a*) para el problema 4.58. [*Sugerencia:* Usar $1 + 2 + 3 \cdots + (n-1) = \frac{1}{2}n(n-1)$, $1^2 + 2^2 + 3^2 + \cdots + (n-1)^2 = \frac{1}{6}n(n-1)(2n-1)$].

4.70 Generalizar y probar la propiedad 3 de este capítulo.

RELACIONES EMPÍRICAS ENTRE LAS MEDIDAS DE DISPERSIÓN

4.71 Comparando las desviaciones estándar obtenidas en el problema 4.59 con las desviaciones medias correspondientes de los problemas 4.41, 4.42 y 4.44, determinar si se cumple la siguiente relación empírica: desviación media $= \frac{4}{5}$(desviación estándar). Explicar cualquier diferencia que se presente.

4.72 Comparando las desviaciones estándar obtenidas en el problema 4.59 con los correspondientes rangos semiintercuartiles del problema 4.48, determinar si se cumple la siguiente relación empírica: rango semiintercuartil $= \frac{2}{3}$(desviación estándar). Explicar cualquier diferencia que se presente.

4.73 ¿Qué relación empírica se espera que exista entre el rango semiintercuartil y la desviación media en distribuciones en forma de campana ligeramente sesgadas?

4.74 Una distribución de frecuencias que es aproximadamente normal tiene un rango semiintercuartil igual a 10. ¿Qué valor se espera que tenga: *a*) la desviación estándar y *b*) la desviación media?

DISPERSIÓN ABSOLUTA Y RELATIVA; COEFICIENTE DE VARIACIÓN

4.75 En un examen final de estadística, la calificación media en un grupo de 150 alumnos es 78 y la desviación estándar 8.0. En álgebra, la puntuación media final del grupo es 73 y la desviación estándar 7.6. ¿En qué materia hay: *a*) mayor dispersión absoluta y *b*) mayor dispersión relativa?

4.76 Encontrar el coeficiente de variación de los datos: *a*) del problema 3.59 y *b*) del problema 3.107.

4.77 En las calificaciones obtenidas por los estudiantes en un examen de admisión, el primer cuartil es 825 y el segundo cuartil es 1 125. Calcular el coeficiente cuartil de variación en estas calificaciones del examen da admisión.

4.78 En el grupo de edad de 15 a 24 años, el primer cuartil de ingreso familiar es $16 500 y el tercer cuartil de ingreso familiar, en este mismo grupo de edad, es $25 000. Calcular el coeficiente cuartil de variación de la distribución de los ingresos en este grupo de edad.

VARIABLES ESTANDARIZADAS; PUNTUACIONES ESTÁNDAR

4.79 En el examen del problema 4.75 la calificación de un estudiante en estadística es 75 y en álgebra 71. ¿En qué examen tiene una puntuación relativa más alta?

4.80 Convertir el conjunto 6, 2, 8, 7, 5 en puntuaciones estándar.

4.81 Probar que la media y la desviación estándar en un conjunto de puntuaciones estándar son iguales a 0 y 1, respectivamente. Emplear el problema 4.80 para ilustrar esto.

4.82 *a*) Convertir las calificaciones del problema 3.107 en puntuaciones estándar y *b*) construir una gráfica de frecuencias relativas contra puntuaciones estándar.

SOFTWARE Y MEDIDAS DE DISPERSIÓN

4.83 En la tabla 4.11 se da el ingreso per cápita en los 50 estados de Estados Unidos, en 2005.

Tabla 4.11 Ingreso per cápita en los 50 estados de Estados Unidos

Estado	Ingreso per cápita	Estado	Ingreso per cápita
Wyoming	36 778	Pennsylvania	34 897
Montana	29 387	Wisconsin	33 565
North Dakota	31 395	Massachusetts	44 289
New Mexico	27 664	Missouri	31 899
West Virginia	27 215	Idaho	28 158
Rhode Island	36 153	Kentucky	28 513
Virginia	38 390	Minnesota	37 373
South Dakota	31 614	Florida	33 219
Alabama	29 136	South Carolina	28 352
Arkansas	26 874	New York	40 507
Maryland	41 760	Indiana	31 276
Iowa	32 315	Connecticut	47 819
Nebraska	33 616	Ohio	32 478
Hawaii	34 539	New Hampshire	38 408
Mississippi	25 318	Texas	32 462
Vermont	33 327	Oregon	32 103
Maine	31 252	New Jersey	43 771
Oklahoma	29 330	California	37 036
Delaware	37 065	Colorado	37 946
Alaska	35 612	North Carolina	30 553
Tennessee	31 107	Illinois	36 120
Kansas	32 836	Michigan	33 116
Arizona	30 267	Washington	35 409
Nevada	35 883	Georgia	31 121
Utah	28 061	Louisiana	24 820

El análisis de estos datos obtenido con SPSS es el siguiente:

Estadística descriptiva

	N	Rango	Desviación estándar	Varianza
Ingresos	50	22 999.00	4 893.54160	2E+007
N validado	50			

Verificar el rango, la desviación estándar y la varianza.

MOMENTOS, SESGO Y CURTOSIS

5

MOMENTOS

Dados N valores $X_1, X_2, \ldots, X_N$ que toma la variable X, se define la cantidad

$$\overline{X^r} = \frac{X_1^r + X_2^r + \cdots + X_N^r}{N} = \frac{\sum_{j=1}^{N} X_j^r}{N} = \frac{\sum X^r}{N} \tag{1}$$

a la que se le llama el r-ésimo *momento*. El primer momento, en el que $r = 1$ es la media aritmética $\bar{X}$.

El r-ésimo *momento respecto a la media* $\bar{X}$ se define como

$$m_r = \frac{\sum_{j=1}^{N} (X_j - \bar{X})^r}{N} = \frac{\sum (X - \bar{X})^r}{N} = \overline{(X - \bar{X})^r} \tag{2}$$

Si $r = 1$, entonces $m_1 = 0$ (ver el problema 3.16). Si $r = 2$, entonces m_2 es la varianza.

El r-ésimo *momento respecto a cualquier origen* A se define de la manera siguiente

$$m'_r = \frac{\sum_{j=1}^{N} (X_j - A)^r}{N} = \frac{\sum (X - A)^r}{N} = \frac{\sum d^r}{N} = \overline{(X - A)^r} \tag{3}$$

donde las $d = X - A$ son las desviaciones de las X respecto de A. Si $A = 0$, la ecuación (*3*) se reduce a la ecuación (*1*). Debido a esto, a la ecuación (*1*) suele llamársele el r-ésimo *momento respecto de cero*.

MOMENTOS PARA DATOS AGRUPADOS

Si $X_1, X_2, \ldots, X_K$ se presentan con frecuencias $f_1, f_2, \ldots, f_K$, respectivamente, los momentos anteriores están dados por

$$\overline{X^r} = \frac{f_1 X_1^r + f_2 X_2^r + \cdots + f_K X_K^r}{N} = \frac{\sum_{j=1}^{K} f_j X_j^r}{N} = \frac{\sum f X^r}{N} \tag{4}$$

$$m_r = \frac{\sum_{j=1}^{K} f_j (X_j - \bar{X})^r}{N} = \frac{\sum f (X - \bar{X})^r}{N} = \overline{(X - \bar{X})^r} \tag{5}$$

$$m'_r = \frac{\sum_{j=1}^{K} f_j (X_j - A)^r}{N} = \frac{\sum f\,(X - A)^r}{N} = \overline{(X - A)^r} \tag{6}$$

donde $N = \sum_{j=1}^{K} f_j = \sum f$. Estas fórmulas se emplean para el cálculo de momentos de datos agrupados.

RELACIONES ENTRE MOMENTOS

Entre los momentos respecto a la media m_r y los momentos respecto de un origen arbitrario m'_r existen las relaciones siguientes:

$$\begin{aligned} m_2 &= m'_2 - m'^2_1 \\ m_3 &= m'_3 - 3m'_1 m'_2 + 2m'^3_1 \\ m_4 &= m'_4 - 4m'_1 m'_3 + 6m'^2_1 m'_2 - 3m'^4_1 \end{aligned} \tag{7}$$

etcétera (ver problema 5.5). Obsérvese que $m'_1 = \bar{X} - A$.

CÁLCULO DE MOMENTOS PARA DATOS AGRUPADOS

El método de compilación dado en capítulos anteriores para el cálculo de la media y de la desviación estándar también puede usarse para obtener un método abreviado para el cálculo de los momentos. Este método aprovecha el hecho de que $X_j = A + cu_j$ (o brevemente, $X = A + cu$), de manera que de acuerdo con la ecuación (*6*) se tiene

$$m'_r = c^r \frac{\sum f u^r}{N} = c^r \overline{u^r} \tag{8}$$

que puede usarse para hallar m_r empleando las ecuaciones (*7*).

COMPROBACIÓN DE CHARLIER Y CORRECCIÓN DE SHEPPARD

La comprobación de Charlier al calcular momentos mediante el método de compilación emplea las identidades:

$$\begin{aligned} \sum f(u+1) &= \sum fu + N \\ \sum f(u+1)^2 &= \sum fu^2 + 2\sum fu + N \\ \sum f(u+1)^3 &= \sum fu^3 + 3\sum fu^2 + 3\sum fu + N \\ \sum f(u+1)^4 &= \sum fu^4 + 4\sum fu^3 + 6\sum fu^2 + 4\sum fu + N \end{aligned} \tag{9}$$

Las correcciones de Sheppard para momentos son las siguientes:

$$m_2 \text{ corregido} = m_2 - \tfrac{1}{12}c^2 \qquad m_4 \text{ corregido} = m_4 - \tfrac{1}{2}c^2 m_2 + \tfrac{7}{240}c^4$$

Los momentos m_1 y m_3 no necesitan corrección.

MOMENTOS EN FORMA ADIMENSIONAL

Para evitar usar unidades particulares, se definen *momentos adimensionales* respecto de la media:

$$a_r = \frac{m_r}{s^r} = \frac{m_r}{(\sqrt{m_2})^r} = \frac{m_r}{\sqrt{m_2^r}} \tag{10}$$

donde $s = \sqrt{m_2}$ es la desviación estándar. Como $m_1 = 0$ y $m_2 = s^2$, se tiene $a_1 = 0$ y $a_2 = 1$.

SESGO

El *sesgo* de una distribución es su grado de asimetría o el grado en el que se aleja de la simetría. Si una curva de frecuencias (polígono de frecuencias suavizado) de una distribución tiene una cola más larga hacia la derecha del máximo central que hacia la izquierda, se dice que la distribución es *sesgada a la derecha*, o que tiene un *sesgo positivo*. Si ocurre lo contrario, se dice que es *sesgada a la izquierda* o que tiene un *sesgo negativo*.

En las distribuciones sesgadas, la media tiende a encontrarse del mismo lado que la cola más larga opuesto al de la moda y que la cola más larga (ver figuras 3-1 y 3-2). Por lo tanto, una medida de la simetría (o sesgo) se obtiene mediante la diferencia: media – moda. Esta medida se puede hacer adimensional dividiendo entre una medida de dispersión, como la desviación estándar, lo que conduce a la definición:

$$\text{Sesgo} = \frac{\text{media} - \text{moda}}{\text{desviación estándar}} = \frac{\bar{X} - \text{moda}}{s} \tag{11}$$

Para evitar el uso de la moda se puede utilizar la fórmula empírica (*10*) del capítulo 3 y definir

$$\text{Sesgo} = \frac{3(\text{media} - \text{mediana})}{\text{desviación estándar}} = \frac{3(\bar{X} - \text{mediana})}{s} \tag{12}$$

A las ecuaciones (*11*) y (*12*) se les llama, respectivamente, *primer coeficiente de sesgo de Pearson* y *segundo coeficiente de sesgo de Pearson*.

Otras medidas del sesgo, que se definen en términos de cuartiles y percentiles, son las siguientes:

$$\text{Coeficiente cuartil de sesgo} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} \tag{13}$$

$$\text{Coeficiente de sesgo percentil 10–90} = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{P_{90} - P_{10}} = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} \tag{14}$$

En una importante medida del sesgo se emplea el tercer momento respecto de la media, tal medida expresada en forma adimensional viene dada por:

$$\text{Coeficiente momento de sesgo} = a_3 = \frac{m_3}{s^3} = \frac{m_3}{(\sqrt{m_2})^3} = \frac{m_3}{\sqrt{m_2^3}} \tag{15}$$

Otra medida de sesgo suele darse mediante $b_1 = a_3^2$. En las curvas perfectamente simétricas, por ejemplo en la curva normal, a_3 y b_1 son cero.

CURTOSIS

La *curtosis* indica qué tan puntiaguda es una distribución; esto por lo regular es en relación con la distribución normal. A una distribución que tiene un pico relativamente alto se le llama *leptocúrtica*, en tanto que si es relativamente aplastada se dice *platicúrtica*. Una distribución normal, que no es ni puntiaguda ni muy aplastada se llama *mesocúrtica*.

En una medida de la curtosis se emplea el cuarto momento respecto de la media, expresada en forma adimensional, esta medida se encuentra dada por:

$$\text{Coeficiente momento de curtosis} = a_4 = \frac{m_4}{s^4} = \frac{m_4}{m_2^2} \tag{16}$$

el cual suele denotarse b_2. En las distribuciones normales $b_2 = a_4 = 3$. A esto se debe que la curtosis suela definirse mediante $(b_2 - 3)$, que tiene signo positivo en una distribución leptocúrtica, negativo en una distribución platicúrtica y cero en las distribuciones normales.

Otra medida de la curtosis se basa tanto en los cuartiles como en los percentiles y está dada por

$$\kappa = \frac{Q}{P_{90} - P_{10}} \tag{17}$$

donde $Q = \frac{1}{2}(Q_3 - Q_1)$ es el rango semiintercuartil. A κ (letra griega minúscula kappa) se le conoce como *coeficiente percentil de curtosis*; en las distribuciones normales, el valor de κ es 0.263.

MOMENTOS, SESGO Y CURTOSIS POBLACIONALES

Cuando es necesario distinguir los momentos muestrales, las medidas de sesgo muestrales o las medidas de curtosis muestrales, de las correspondientes medidas de la población de la que es parte la muestra, se acostumbra usar letras del alfabeto latino para las primeras y letras del alfabeto griego para las últimas. Así, si los momentos muestrales se denotan m_r y m'_r, los correspondientes momentos poblacionales serán, μ_r y μ'_r (μ es la letra *mu* del alfabeto griego). Como subíndices se emplean siempre letras del alfabeto latino.

De igual manera, si las medidas muestrales de sesgo y curtosis se denotan a_3 y a_4, respectivamente, los sesgos y las curtosis poblacionales serán α_3 y α_4 (α es la letra *alfa* del alfabeto griego).

Como ya se dijo en el capítulo 4, la desviación estándar de una muestra y la desviación estándar de una población se denotan s y σ, respectivamente.

CÁLCULO DEL SESGO Y DE LA CURTOSIS EMPLEANDO SOFTWARE

El software visto en este libro puede usarse para calcular las medidas de curtosis y de sesgo de datos muestrales. Los datos que se presentan en la tabla 5.1 son muestras de 50 elementos (de tamaño 50) tomadas de distribuciones, una normal, otra sesgada a la derecha, otra sesgada a la izquierda y la última es una distribución uniforme.

Los datos normales son estaturas de mujeres, los datos sesgados a la derecha son edades de casamiento de mujeres, los datos sesgados a la izquierda son edades a las que fallecen las mujeres, y los datos uniformes son cantidades de

Tabla 5.1

Normal		Sesgada a la derecha		Sesgada a la izquierda		Uniforme	
67	69	31	40	102	87	12.1	11.6
70	62	43	24	55	104	12.1	11.6
63	67	30	29	70	75	12.4	12.0
65	59	30	24	95	80	12.1	11.6
68	66	38	27	73	66	12.1	11.6
60	65	26	35	79	93	12.2	11.7
70	63	29	33	60	90	12.2	12.3
64	65	55	75	73	84	12.2	11.7
69	60	46	38	89	73	11.9	11.7
61	67	26	34	85	98	12.2	11.7
66	64	29	85	72	79	12.3	11.8
65	68	57	29	92	35	12.3	12.5
71	61	34	40	76	71	11.7	11.8
62	69	34	41	93	90	12.3	11.8
66	65	36	35	76	71	12.3	11.8
68	62	40	26	97	63	12.4	11.9
64	67	28	34	10	58	12.4	11.9
67	70	26	19	70	82	12.1	11.9
62	64	66	23	85	72	12.4	12.2
66	63	63	28	25	93	12.4	11.9
65	68	30	26	83	44	12.5	12.0
63	64	33	31	58	65	11.8	11.9
66	65	24	25	10	77	12.5	12.0
65	61	35	22	92	81	12.5	12.0
63	66	34	28	82	77	12.5	12.0

refresco despachadas por una máquina en envases de 12 onzas. En la figura 5-1 se muestra la distribución de cada uno de estos conjuntos de datos muestrales. Las distribuciones de las cuatro muestras se ilustran mediante gráficas de puntos.

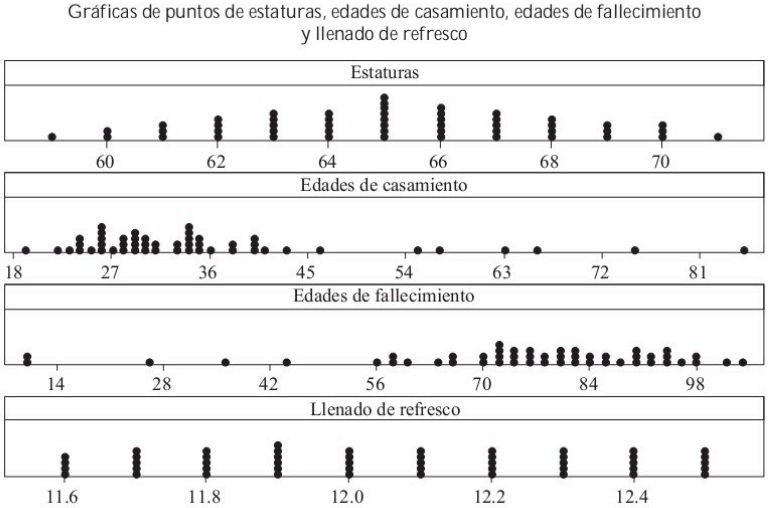


Figura 5-1 MINITAB, gráficas de cuatro distribuciones: normal, sesgada a la derecha, sesgada a la izquierda y uniforme.

En la variable estatura se da la estatura de 50 mujeres adultas, en la variable edad de casamiento se da la edad de casamiento de 50 mujeres, en la variable edad de fallecimiento se da la edad de fallecimiento de 50 mujeres y en la variable llenado de refresco se dan las cantidades de refresco despachadas en recipientes de 12 onzas. Cada muestra tiene 50 elementos (es de tamaño 50). Empleando la terminología aprendida en este capítulo: la distribución de las estaturas es mesocúrtica, la distribución del llenado de refresco es platicúrtica, la distribución de las edades de casamiento es sesgada a la derecha, y la distribución de las edades de fallecimiento es sesgada a la izquierda.

EJEMPLO 1 Para hallar los valores correspondientes al sesgo y a la curtosis de las cuatro variables puede emplearse MINITAB. Seleccionando la secuencia "**Stat ⇒ Basic statistics ⇒ Display descriptive statistics**", se obtiene el siguiente resultado:

Estadísticos descritivos: estatura, edades de casamiento, edades de fallecimiento y llenado de refresco

Variable	N	N*	Mean	StDev	Skewness	Kurtosis
Height	50	0	65.120	2.911	−0.02	−0.61
Wedage	50	0	35.48	13.51	1.98	4.10
Obitage	50	0	74.20	20.70	−1.50	2.64
Cola-fill	50	0	12.056	0.284	0.02	−1.19

Como se ve, los valores dados para el sesgo de las distribuciones normal y uniforme son cercanos a 0; el sesgo es positivo para la distribución sesgada a la derecha y negativo para la distribución sesgada a la izquierda.

EJEMPLO 2 Use EXCEL para hallar los valores de sesgo y curtosis correspondientes a los datos de la figura 5-1. Los nombres de las variables se ingresan en A1:D1, los datos muestrales se ingresan en A2:D51 y en cualquier celda vacía se ingresa =COEFICIENTE.ASIMETRIA(A2:A51) obteniéndose como resultado −0.0203. La función =COEFICIENTE.ASIMETRIA(B2:B51) da como resultado 1.9774, la función =COEFICIENTE.ASIMETRIA(C2:C51) da como resultado −1.4986 y la función =COEFICIENTE.ASIMETRIA(D2:D51) da como resultado 0.0156. Los valores de curtosis se obtienen mediante las funciones =CURTOSIS(A2:A51) que da como resultado −0.6083, =CURTOSIS(B2:B51) que da como resultado 4.0985, =CURTOSIS(C2:C51) que da como resultado 2.6368 y =CURTOSIS(D2:D51) que da como resultado −1.1889. Como puede observarse, EXCEL y MINITAB dan los mismos valores de curtosis y de sesgo.

EJEMPLO 3 También se puede emplear STATISTIX para analizar los datos de la figura 5-1. Se selecciona la secuencia "**Statistics ⇒ Summary Statistics ⇒ Descriptive Statistics**" y se obtiene la ventana de diálogo que se muestra en la figura 5-2.

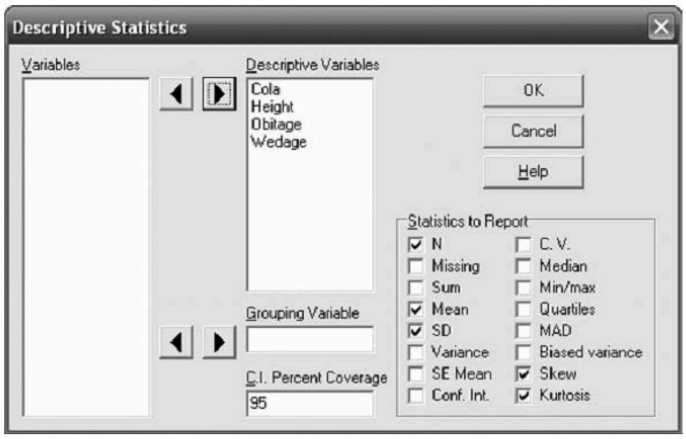


Figura 5-2 Ventana de diálogo de STATISTIX.

Obsérvese que N, Mean (media), SD (desviación estándar), Skew (sesgo) y Kurtosis (curtosis) fueron seleccionados como los estadísticos que se desean conocer. El resultado que se obtiene de STATISTIX es:

Estadísticos descriptivos

Variable	N	Mean	SD	Skew	Kurtosis
Cola	50	12.056	0.2837	0.0151	−1.1910
Height	50	65.120	2.9112	−0.0197	−0.6668
Obitage	50	74.200	20.696	−1.4533	2.2628
Wedage	50	35.480	13.511	1.9176	3.5823

Como los valores numéricos difieren ligeramente de los obtenidos con EXCEL y MINITAB, es claro que este software emplea métodos ligeramente diferentes para medir la curtosis y el sesgo.

EJEMPLO 4 En SPSS con la secuencia "**Analyze ⇒ Descriptive Statistics ⇒ Descriptives**" se obtiene la ventana de diálogo que se presenta en la figura 5-3, en la cual se selecciona Mean (media), Std. deviation (desviación estándar), Kurtosis (curtosis) y Skewness (sesgo). SPSS da las mismas medidas de sesgo y curtosis que EXCEL y MINITAB.

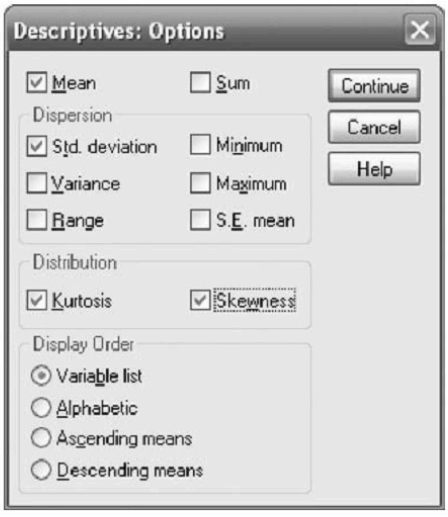


Figura 5-3 Ventana de diálogo de SPSS.

SPSS proporciona los siguientes resultados.

Estadísticos descriptivos

	N	Media	Desv. Estándar	Sesgo		Curtosis	
	Estadístico	Estadístico	Estadístico	Estadístico	Error estándar	Estadístico	Error estándar
Estatura	50	65.1200	2.91120	−.020	.337	−.608	.662
Casamientos	50	35.4800	13.51075	1.977	.337	4.098	.662
Defunciones	50	74.2000	20.69605	−1.499	.337	2.637	.662
Llenado	50	12.0560	.28368	.016	.337	−1.189	.662
N validada	50						

EJEMPLO 5 Si se usa SAS para calcular los valores del sesgo y de la curtosis, se obtienen los resultados que se muestran a continuación. Estos resultados son prácticamente los mismos que se obtienen con EXCEL, MINITAB y SPSS.

```
                        The MEANS Procedure
  Variable      Mean          Std Dev       N    Skewness      Kurtosis
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
  Height       65.1200000    2.9112029    50   −0.0203232   −0.6083437
  Wedage       35.4800000   13.5107516    50    1.9774237    4.0984607
  Obitage      74.2000000   20.6960511    50   −1.4986145    2.6368045
  Cola_fill    12.0560000    0.2836785    50    0.0156088   −1.1889600
ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
```

PROBLEMAS RESUELTOS

MOMENTOS

5.1 Encontrar: *a*) el primero, *b*) el segundo, *c*) el tercero y *d*) el cuarto momentos del conjunto 2, 3, 7, 8, 10.

SOLUCIÓN

a) El primer momento o media aritmética es

$$\bar{X} = \frac{\sum X}{N} = \frac{2+3+7+8+10}{5} = \frac{30}{5} = 6$$

b) El segundo momento es

$$\overline{X^2} = \frac{\sum X^2}{N} = \frac{2^2+3^2+7^2+8^2+10^2}{5} = \frac{226}{5} = 45.2$$

c) El tercer momento es

$$\overline{X^3} = \frac{\sum X^3}{N} = \frac{2^3+3^3+7^3+8^3+10^3}{5} = \frac{1\,890}{5} = 378$$

d) El cuarto momento es

$$\overline{X^4} = \frac{\sum X^4}{N} = \frac{2^4+3^4+7^4+8^4+10^4}{5} = \frac{16\,594}{5} = 3\,318.8$$

5.2 Dado el conjunto de números del problema 5.1, encontrar: *a*) el primero, *b*) el segundo, *c*) el tercero y *d*) el cuarto momentos respecto de la media.

SOLUCIÓN

a) $m_1 = \overline{(X - \bar{X})} = \dfrac{\sum (X - \bar{X})}{N} = \dfrac{(2-6) + (3-6) + (7-6) + (8-6) + (10-6)}{5} = \dfrac{0}{5} = 0$

m_1 siempre es igual a cero debido a que $\overline{X - \bar{X}} = \bar{X} - \bar{X} = 0$ (ver problema 3.16).

b) $m_2 = \overline{(X - \bar{X})^2} = \dfrac{\sum (X - \bar{X})^2}{N} = \dfrac{(2-6)^2 + (3-6)^2 + (7-6)^2 + (8-6)^2 + (10-6)^2}{6} = \dfrac{46}{5} = 9.2$

Obsérvese que m_2 es la varianza s^2.

c) $m_3 = \overline{(X - \bar{X})^3} = \dfrac{\sum (X - \bar{X})^3}{N} = \dfrac{(2-6)^3 + (3-6)^3 + (7-6)^3 + (8-6)^3 + (10-6)^3}{5} = \dfrac{-18}{5} = -3.6$

d) $m_4 = \overline{(X - \bar{X})^4} = \dfrac{\sum (X - \bar{X})^4}{N} = \dfrac{(2-6)^4 + (3-6)^4 + (7-6)^4 + (8-6)^4 + (10-6)^4}{5} = \dfrac{610}{5} = 122$

5.3 Para el conjunto de números del problema 5.1, encontrar: *a*) el primero, *b*) el segundo, *c*) el tercero y *d*) el cuarto momentos respecto del origen.

SOLUCIÓN

a) $m_1' = \overline{(X - 4)} = \dfrac{\sum (X - 4)}{N} = \dfrac{(2-4) + (3-4) + (7-4) + (8-4) + (10-4)}{5} = 2$

b) $m_2' = \overline{(X - 4)^2} = \dfrac{\sum (X - 4)^2}{N} = \dfrac{(2-4)^2 + (3-4)^2 + (7-4)^2 + (8-4)^2 + (10-4)^2}{5} = \dfrac{66}{5} = 13.2$

c) $m_3' = \overline{(X - 4)^3} = \dfrac{\sum (X - 4)^3}{N} = \dfrac{(2-4)^3 + (3-4)^3 + (7-4)^3 + (8-4)^3 + (10-4)^3}{5} = \dfrac{298}{5} = 59.6$

d) $m_4' = \overline{(X - 4)^4} = \dfrac{\sum (X - 4)^4}{N} = \dfrac{(2-4)^4 + (3-4)^4 + (7-4)^4 + (8-4)^4 + (10-4)^4}{5} = \dfrac{1\,650}{5} = 330$

5.4 Empleando los resultados de los problemas 5.2 y 5.3, verificar las siguientes relaciones entre los momentos: *a*) $m_2 = m_2' - m_1'^2$, *b*) $m_3 = m_3' - 3m_1'm_2' + 2m_1'^3$ y *c*) $m_4 = m_4' - 4m_1'm_3' + 6m_1'^2m_2' - 3m_1'^4$.

SOLUCIÓN

De acuerdo con el problema 5.3 se tiene $m_1' = 2$, $m_2' = 13.2$, $m_3' = 59.6$ y $m_4' = 330$. Por lo tanto:

a) $m_2 = m_2' - m_1'^2 = 13.2 - (2)^2 = 13.2 - 4 = 9.2$

b) $m_3 = m_3' - 3m_1'm_2' + 2m_1'^3 = 59.6 - (3)(2)(13.2) + 2(2)^3 = 59.6 - 79.2 + 16 = -3.6$

c) $m_4 = m_4' - 4m_1'm_3' + 6m_1'^2m_2' - 3m_1'^4 = 330 - 4(2)(59.6) + 6(2)^2(13.2) - 3(2)^4 = 122$

lo que coincide con el problema 5.2.

5.5 Probar que: *a*) $m_2 = m_2' - m_1'^2$, *b*) $m_3 = m_3' - 3m_1'm_2' + 2m_1'^3$ y *c*) $m_4 = m_4' - 4m_1'm_3' + 6m_1'^2m_2' - 3m_1'^4$.

SOLUCIÓN

Si $d = X - A$, entonces $X = A + d$, $\bar{X} = A + \bar{d}$ y $X - \bar{X} = d - \bar{d}$, y por lo tanto:

a)
$$m_2 = \overline{(X - \bar{X})^2} = \overline{(d - \bar{d})^2} = \overline{d^2 - 2\bar{d}d + \bar{d}^2}$$
$$= \overline{d^2} - 2\bar{d}^2 + \bar{d}^2 = \overline{d^2} - \bar{d}^2 = m_2' - m_1'^2$$

b) $m_3 = \overline{(X-\bar{X})^3} = \overline{(d-\bar{d})^3} = \overline{(d^3 - 3d^2\bar{d} + 3d\bar{d}^2 - \bar{d}^3)}$
$= \overline{d^3} - 3\bar{d}\overline{d^2} + 3\bar{d}^3 - \bar{d}^3 = \overline{d^3} - 3\bar{d}\overline{d^2} + 2\bar{d}^3 = m_3' - 3m_1'm_2' + 2m_1'^3$

c) $m_4 = \overline{(X-\bar{X})^4} = \overline{(d-\bar{d})^4} = \overline{(d^4 - 4d^3\bar{d} + 6d^2\bar{d}^2 - 4d\bar{d}^3 + \bar{d}^4)}$
$= \overline{d^4} - 4\bar{d}\overline{d^3} + 6\bar{d}^2\overline{d^2} - 4\bar{d}^4 + \bar{d}^4 = \overline{d^4} - 4\bar{d}\overline{d^3} + 6\bar{d}^2\overline{d^2} - 3\bar{d}^4$
$= m_4' - 4m_1'm_3' + 6m_1'^2m_2' - 3m_1'^4$

Por extensión de este método se pueden deducir fórmulas semejantes para m_5, m_6, etcétera.

CÁLCULO DE MOMENTOS PARA DATOS AGRUPADOS

5.6 Encuentre los primeros cuatro momentos respecto de la media para la distribución de estaturas del problema 3.22.

SOLUCIÓN

Para facilitar los cálculos se pueden disponer como en la tabla 5.2, a partir de la cual se tiene

$$m_1' = c\frac{\sum fu}{N} = (3)\left(\frac{15}{100}\right) = 0.45 \qquad m_3' = c^3\frac{\sum fu^3}{N} = (3)^3\left(\frac{33}{100}\right) = 8.91$$

$$m_2' = c^2\frac{\sum fu^2}{N} = (3)^2\left(\frac{97}{100}\right) = 8.73 \qquad m_4' = c^4\frac{\sum fu^4}{N} = (3)^4\left(\frac{253}{100}\right) = 204.93$$

Por lo tanto
$$m_1 = 0$$
$$m_2 = m_2' - m_1'^2 = 8.73 - (0.45)^2 = 8.5275$$
$$m_3 = m_3' - 3m_1'm_2' + m_1'^3 = 8.91 - 3(0.45)(8.73) + 2(0.45)^3 = -2.6932$$
$$m_4 = m_4' - 4m_1'm_3' + 6m_1'^2m_2' - 3m_1'^4$$
$$= 204.93 - 4(0.45)(8.91) + 6(0.45)^2(8.73) - 3(0.45)^4 = 199.3759$$

Tabla 5.2

X	u	f	fu	fu^2	fu^3	fu^4
61	−2	5	−10	20	−40	80
64	−1	18	−18	18	−18	18
67	0	42	0	0	0	0
70	1	27	27	27	27	27
73	2	8	16	32	64	128
		$N = \sum f = 10$	$\sum fu = 15$	$\sum fu^2 = 97$	$\sum fu^3 = 33$	$\sum fu^4 = 253$

5.7 Encontrar: *a)* m_1', *b)* m_2', *c)* m_3', *d)* m_4', *e)* m_1, *f)* m_2, *g)* m_3, *h)* m_4, *i)* $\bar{X}$, *j)* s, *k)* $\overline{X^2}$ y *l)* $\overline{X^3}$ para la distribución de la tabla 4.7 del problema 4.19.

SOLUCIÓN

Para facilitar los cálculos, disponerlos como en la tabla 5.3.

Tabla 5.3

	X	u	f	fu	fu^2	fu^3	fu^4
	70	−6	4	−24	144	−864	5 184
	74	−5	9	−45	225	−1 125	5 625
	78	−4	16	−64	256	−1 024	4 096
	82	−3	28	−84	252	−756	2 268
	86	−2	45	−90	180	−360	720
	90	−1	66	−66	66	−66	66
$A \longrightarrow$	94	0	85	0	0	0	0
	98	1	72	72	72	72	72
	102	2	54	108	216	432	864
	106	3	38	114	342	1 026	3 078
	110	4	27	108	432	1 728	6 912
	114	5	18	90	450	2 250	11 250
	118	6	11	66	396	2 376	14 256
	122	7	5	34	245	1 715	12 005
	126	8	2	16	128	1 024	8 192
			$N = \sum f = 480$	$\sum fu = 236$	$\sum fu^2 = 3\,404$	$\sum fu^3 = 6\,428$	$\sum fu^4 = 74\,588$

a) $m_1' = c\dfrac{\sum fu}{N} = (4)\left(\dfrac{236}{480}\right) = 1.9667$

b) $m_2' = c^2\dfrac{\sum fu^2}{N} = (4)^2\left(\dfrac{3\,404}{480}\right) = 113.4667$

c) $m_3' = c^3\dfrac{\sum fu^3}{N} = (4)^3\left(\dfrac{6\,428}{480}\right) = 857.0667$

d) $m_4' = c^4\dfrac{\sum fu^4}{N} = (4)^4\left(\dfrac{74\,588}{480}\right) = 39\,780.2667$

e) $m_1 = 0$

f) $m_2 = m_2' - m_1'^2 = 113.4667 - (1.9667)^2 = 109.5988$

g) $m_3 = m_3' - 3m_1'm_2' + 2m_1'^3 = 857.0667 - 3(1.9667)(113.4667) + 2(1.9667)^3 = 202.8158$

h) $m_4 = m_4' - 4m_1'm_3' + 6m_1'^2m_2' - 3m_1'^4 = 35\,627.2853$

i) $\bar{X} = \overline{(A+d)} = A + m_1' = A + c\dfrac{\sum fu}{N} = 94 + 1.9667 = 95.97$

j) $s = \sqrt{m_2} = \sqrt{109.5988} = 10.47$

k) $\overline{X^2} = \overline{(A+d)^2} = \overline{(A^2 + 2Ad + d^2)} = A^2 + 2A\bar{d} + \overline{d^2} = A^2 + 2Am_1' + m_2'$
$= (94)^2 + 2(94)(1.9667) + 113.4667 = 9\,319.2063$ o 9 319 a cuatro cifras significativas.

l) $\overline{X^3} = \overline{(A+d)^3} = \overline{(A^3 + 3A^2d + 3Ad^2 + d^3)} = A^3 + 3A^2\bar{d} + 3A\overline{d^2} + \overline{d^3}$
$= A^3 + 3A^2m_1' + 3Am_2' + m_3' = 915\,571.9597$ o 915 600 a cuatro cifras significativas.

COMPROBACIÓN DE CHARLIER

5.8 Ilustrar el uso de la comprobación de Charlier en los cálculos del problema 5.7.

SOLUCIÓN

Para proporcionar la comprobación deseada, a la tabla 5.3 se agregan las columnas de la tabla 5.4 (con excepción de la columna 2, que por comodidad se repite en la tabla 5.3).

En cada uno de los siguientes pares de fórmulas, los datos para la primera se toman de la tabla 5.4 y los datos para la segunda se toman de la tabla 5.2. La igualdad de los resultados en cada par proporciona la comprobación deseada.

Tabla 5.4

$u+1$	f	$f(u+1)$	$f(u+1)^2$	$f(u+1)^3$	$f(u+1)^4$
−5	4	−20	100	−500	2 500
−4	9	−36	144	−576	2 304
−3	16	−48	144	−432	1 296
−2	28	−56	112	−224	448
−1	45	−45	45	−45	45
0	66	0	0	0	0
1	85	85	85	85	85
2	72	144	288	576	1 152
3	54	162	486	1 458	4 374
4	38	152	608	2 432	9 728
5	27	135	675	3 375	16 875
6	18	108	648	3 888	23 328
7	11	77	539	3 773	26 411
8	5	40	320	2 560	20 480
9	2	18	162	1 458	13 122
	$N=\sum f$ = 480	$\sum f(u+1)$ = 716	$\sum f(u+1)^2$ = 4 356	$\sum f(u+1)^3$ = 17 828	$\sum f(u+1)^4$ = 122 148

$\sum f(u+1) = 716$

$\sum fu + N = 236 + 480 = 716$

$\sum f(u+1)^2 = 4\,356$

$\sum fu^2 + 2\sum fu + N = 3\,404 + 2\,(236) + 480 = 4\,356$

$\sum f(u+1)^3 = 17\,828$

$\sum fu^3 + 3\sum fu^2 + 3\sum fu + N = 6\,428 + 3(3\,404) + 3(236) + 480 = 17\,828$

$\sum f(u+1)^4 = 122\,148$

$\sum fu^4 + 4\sum fu^3 + 6\sum fu^2 + 4\sum fu + N = 74\,588 + 4(6\,428) + 6(3\,404) + 4(236) + 480 = 122\,148$

CORRECCIONES DE SHEPPARD PARA LOS MOMENTOS

5.9 Emplear las correcciones de Sheppard para determinar los momentos respecto de la media para los datos: *a*) del problema 5.6 y *b*) del problema 5.7.

SOLUCIÓN

a) m_2 corregido $= m_2 - c^2/12 = 8.5275 - 3^2/12 = 7.7775$

m_4 corregido $= m_4 - \frac{1}{2}c^2 m_2 + \frac{7}{240}c^4$

$= 199.3759 - \frac{1}{2}(3)^2(8.5275) + \frac{7}{240}(3)^4$

$= 163.3646$

m_1 y m_3 no necesitan correcciones.

b) m_2 corregido $= m_2 - c^2/12 = 109.5988 - 4^2/12 = 108.2655$

m_4 corregido $= m_4 - \frac{1}{2}c^2 m_2 + \frac{7}{240}c^4$

$= 35\,627.2853 - \frac{1}{2}(4)^2(109.5988) + \frac{7}{240}(4)^4$

$= 34\,757.9616$

SESGO

5.10 Encontrar: *a*) el primer coeficiente de sesgo de Pearson y *b*) el segundo coeficiente de sesgo de Pearson para la distribución de los salarios de los 65 empleados de la empresa P&R (ver los problemas 3.44 y 4.18).

SOLUCIÓN

Media = \$279.76, mediana = \$279.06, moda = \$277.50 y desviación estándar s = \$15.60. Por lo tanto:

a) $$\text{Primer coeficiente de sesgo} = \frac{\text{media} - \text{moda}}{s} = \frac{\$279.76 - \$277.50}{\$15.60} = 0.1448;\ \text{o bien } 0.14$$

b) $$\text{Segundo coeficiente de sesgo} = \frac{3(\text{media} - \text{mediana})}{s} = \frac{3(\$279.76 - \$279.06)}{\$15.60} = 0.1346;\ \text{o bien } 0.13$$

Si se emplea la desviación estándar corregida [ver problema 4.21*b*)], estos coeficientes se convierten, respectivamente, en:

a) $$\frac{\text{Media} - \text{moda}}{s\ \text{corregida}} = \frac{\$279.76 - \$277.50}{\$15.33} = 0.1474;\ \text{o bien } 0.15$$

b) $$\frac{3(\text{media} - \text{mediana})}{s\ \text{corregida}} = \frac{3(\$279.76 - \$279.06)}{\$15.33} = 0.1370;\ \text{o bien } 0.14$$

Como los coeficientes son positivos, la distribución tiene sesgo positivo (es decir, a la derecha).

5.11 Encontrar los coeficientes: *a*) cuartil y *b*) percentil de sesgo para la distribución del problema 5.10 (ver problema 3.44).

SOLUCIÓN

Q_1 = \$268.25, $Q_2 = P_{50}$ = \$279.06, Q_3 = \$290.75, $P_{10} = D_1$ = \$258.12, y $P_{90} = D_9$ = \$301.00. Por lo tanto:

a) $$\text{Coeficiente cuartil de sesgo} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} = \frac{\$290.75 - 2(\$279.06) + \$268.25}{\$290.75 - \$268.25} = 0.0391$$

b) $$\text{Coeficiente percentil de sesgo} = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} = \frac{\$301.00 - 2(\$279.06) + \$258.12}{\$301.00 - \$258.12} = 0.0233$$

5.12 Encontrar el coeficiente momento de sesgo, a_3, *a*) en la distribución de las estaturas de los estudiantes de la universidad XYZ (ver problema 5.6) y *b*) en los CI de los niños de primaria (ver problema 5.7).

SOLUCIÓN

a) $m_2 = s^2 = 8.5275$ y $m_3 = -2.6932$. Por lo tanto:

$$a_3 = \frac{m_3}{s^3} = \frac{m_3}{(\sqrt{m_2})^3} = \frac{-2.6932}{(\sqrt{8.5275})^3} = -0.1081 \qquad \text{o bien} \qquad -0.11$$

Si se emplea la corrección de Sheppard para datos agrupados [ver problema 5.9*a*)], se tiene

$$a_3 \text{ corregido} = \frac{m_3}{(\sqrt{m_2 \text{ corregido}})^3} = \frac{-2.6932}{(\sqrt{7.7775})^3} = -0.1242 \qquad \text{o bien} \qquad -0.12$$

b) $$a_3 = \frac{m_3}{s^3} = \frac{m_3}{(\sqrt{m_2})^3} = \frac{202.8158}{(\sqrt{109.5988})^3} = 0.1768 \qquad \text{o bien} \qquad 0.18$$

Si se emplea la corrección de Sheppard para datos agrupados [ver problema 5.9*b*)], se tiene

$$a_3 \text{ corregido} = \frac{m_3}{(\sqrt{m_2 \text{ corregido}})^3} = \frac{202.8158}{(\sqrt{108.2655})^3} = 0.1800 \qquad \text{o} \qquad 0.18$$

Obsérvese que ambas distribuciones son moderadamente sesgadas, la distribución *a*) a la izquierda (negativamente) y la distribución *b*) a la derecha (positivamente). La distribución *b*) es más sesgada que la distribución *a*); es decir, *a*) es más simétrica que *b*), lo que es evidente dado que el valor numérico (o el valor absoluto) del coeficiente de sesgo para *b*) es mayor que el valor del coeficiente de sesgo para *a*).

CURTOSIS

5.13 Encontrar el coeficiente momento de curtosis, a_4, de los datos: *a*) del problema 5.6 y *b*) del problema 5.7.

SOLUCIÓN

a) $$a_4 = \frac{m_4}{s^4} = \frac{m_4}{m_2^2} = \frac{199.3759}{(8.5275)^2} = 2.7418 \qquad \text{o bien} \qquad 2.74$$

Si se emplean las correcciones de Sheppard [ver problema 5.9*a*)], entonces

$$a_4 \text{ corregido} = \frac{m_4 \text{ corregido}}{(m_2 \text{ corregido})^2} = \frac{163.36346}{(7.7775)^5} = 2.7007 \qquad \text{o bien} \qquad 2.70$$

b) $$a_4 = \frac{m_4}{s^4} = \frac{m_4}{m_2^2} = \frac{35\,627.2853}{(109.5988)^2} = 2.9660 \qquad \text{o bien} \qquad 2.97$$

Si se emplean las correcciones de Sheppard [ver problema 5.9*b*)], entonces

$$a_4 \text{ corregido} = \frac{m_4 \text{ corregido}}{(m_2 \text{ corregido})^2} = \frac{34\,757.9616}{(108.2655)^2} = 2.9653 \qquad \text{o bien} \qquad 2.97$$

Como en una distribución normal $a_4 = 3$, se sigue que ambas distribuciones, *a*) y *b*), son *platicúrticas* con respecto a la distribución normal (es decir, menos puntiagudas que la distribución normal).

En lo que se refiere a qué tan puntiagudas son, la distribución *b*) se aproxima a la distribución normal más que la distribución *a*). Sin embargo, de acuerdo con el problema 5.12, la distribución *a*) es más simétrica que la distribución *b*), de manera que en lo que se refiere a la simetría, *a*) se aproxima más a una distribución normal que *b*).

CÁLCULO DEL SESGO Y DE LA CURTOSIS EMPLEANDO SOFTWARE

5.14 Algunas veces las puntuaciones de un examen no siguen una distribución normal, aunque generalmente lo hacen. Algunas veces se observa que los estudiantes obtienen puntuaciones altas o bajas y que hay pocas puntuaciones intermedias. La distribución que se muestra en la figura 5-4 es una de estas distribuciones. Este tipo de distribuciones se conocen como distribuciones en forma de U. Empleando EXCEL, encontrar la media, la desviación estándar, el sesgo y la curtosis de estos datos.

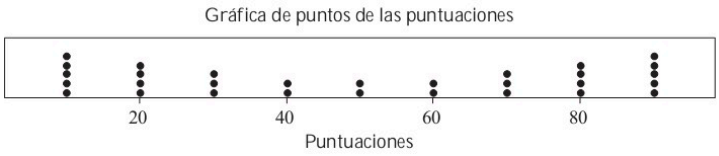


Figura 5-4 MINITAB, gráfica de puntos con datos que siguen una distribución en forma de U.

SOLUCIÓN

En una hoja de cálculo de EXCEL se ingresan los datos en A1:A30. El comando "=**AVERAGE(A1:A30)**" da como resultado 50. El comando "=**STDEV(A1:A30)**" da como resultado 29.94. El comando " =**SKEW(A1:A30)**" da como resultado 0. El comando " =**KURT(A1:A30)**" da como resultado –1.59.

PROBLEMAS SUPLEMENTARIOS

MOMENTOS

5.15 Encontrar el: *a*) primero, *b*) segundo, *c*) tercero y *d*) cuarto momentos del conjunto 4, 7, 5, 9, 8, 3, 6.

5.16 Encontrar el: *a*) primero, *b*) segundo, *c*) tercero y *d*) cuarto momentos respecto de la media para el conjunto de datos del problema 5.15.

5.17 Encontrar el: *a*) primero, *b*) segundo, *c*) tercero y *d*) cuarto momentos respecto del número 7 para el conjunto de datos del problema 5.15.

5.18 Empleando los resultados de los problemas 5.15 y 5.17, verificar las siguientes relaciones entre los momentos: *a*) $m_2 = m_2' - m_1'^2$, *b*) $m_3 = m_3' - 3m_1'm_2' + 2m_1'^3$ y *c*) $m_4 = m_4' - 4m_1'm_3' + 6m_1'^2m_2' - 3m_1'^4$.

5.19 En el conjunto de números de la progresión aritmética 2, 5, 8, 11, 14, 17, encontrar los primeros cuatro momentos respecto de la media.

5.20 Probar que: *a*) $m_2' = m_2 + h^2$, *b*) $m_3' = m_3 + 3hm_2 + h^3$ y *c*) $m_4' = m_4 + 4hm_3 + 6h^2m_2 + h^4$, donde $h = m_1'$.

5.21 Si el primer momento respecto del número 2 es igual a 5, ¿cuál es la media?

5.22 Si los primeros cuatro momentos respecto del número 3 son –2, 10, –25 y 50, determinar los momentos correspondientes: *a*) respecto de la media, *b*) respecto del número 5 y *c*) respecto del cero.

5.23 Para los números 0, 0, 0, 1, 1, 1, 1 y 1, encontrar los primeros cuatro momentos respecto de la media.

5.24 *a*) Probar que $m_5 = m_5' - 5m_1'm_4' + 10m_1'^2m_3' - 10m_1'^3m_2' + 4m_1'^5$.
b) Deducir una fórmula similar para m_6.

5.25 De un total de N números, la fracción p son unos y la fracción $q = 1 - p$ son ceros. Encontrar: *a*) m_1, *b*) m_2, *c*) m_3 y *d*) m_4 para este conjunto de números. Comparar con el problema 5.23.

5.26 Probar que los primeros cuatro momentos respecto de la media en la progresión geométrica $a, a + d, a + 2d, \ldots, a + (n-1)d$ son $m_1 = 0$, $m_2 = \frac{1}{12}(n^2 - 1)d^2$, $m_3 = 0$ y $m_4 = \frac{1}{240}(n^2 - 1)(3n^2 - 7)d^4$. Comparar con el problema 5.19 (ver también el problema 4.69). [*Sugerencia*: $1^4 + 2^4 + 3^4 + \cdots + (n-1)^4 = \frac{1}{30}n(n-1)(2n-1)(3n^2 - 3n - 1)$.]

MOMENTOS PARA DATOS AGRUPADOS

5.27 Dada la distribución de la tabla 5.5, calcular los cuatro momentos respecto de la media.

Tabla 5.5

X	f
12	1
14	4
16	6
18	10
20	7
22	2
	Total 30

5.28 Ilustrar el uso de la comprobación de Charlier para los cálculos del problema 5.27.

5.29 Aplicar las correcciones de Sheppard a los momentos obtenidos en el problema 5.27.

5.30 Dada la distribución del problema 3.59, calcular los primeros cuatro momentos respecto de la media: *a*) sin correcciones de Sheppard y *b*) con correcciones de Sheppard.

5.31 Dada la distribución del problema 3.62, encontrar *a*) m_1, *b*) m_2, *c*) m_3, *d*) m_4, *e*) $\bar{X}$, *f*) s, *g*) $\overline{X^2}$, *h*) $\overline{X^3}$, *i*) $\overline{X^4}$ y *j*) $\overline{(X+1)^3}$.

SESGO

5.32 Encontrar el coeficiente momento de sesgo, a_3, para la distribución del problema 5.27: *a*) sin correcciones y *b*) con correcciones de Sheppard.

5.33 Encontrar el coeficiente momento de sesgo, a_3, para la distribución del problema 3.59 (ver problema 5.30).

5.34 Los segundos momentos respecto de la media de dos distribuciones son 9 y 16, en tanto que los terceros momentos respecto de la media son –8.1 y –12.8, respectivamente. ¿Qué distribución es más sesgada a la izquierda?

5.35 Encontrar los coeficientes de Pearson: *a*) primero y *b*) segundo para la distribución del problema 3.59 y explicar cualquier diferencia que se encuentre.

5.36 Encontrar los coeficientes de sesgo: *a*) cuartil y *b*) percentil para la distribución del problema 3.59. Comparar sus resultados con los del problema 5.35 y explicar.

5.37 En la tabla 5.6 se dan tres distribuciones diferentes de la variable X. Las frecuencias de cada una de las tres distribuciones están dadas por f_1, f_2 y f_3. Encontrar el primero y el segundo coeficientes de sesgo de Pearson de las tres distribuciones. Para calcular los coeficientes, emplear la desviación estándar corregida.

Tabla 5.6

X	f_1	f_2	f_3
0	10	1	1
1	5	2	2
2	2	14	2
3	2	2	5
4	1	1	10

CURTOSIS

5.38 Encontrar el coeficiente momento de curtosis, a_4, para la distribución del problema 5.27: *a*) sin correcciones y *b*) con correcciones de Sheppard.

5.39 Encontrar el coeficiente momento de curtosis, a_4, de la distribución del problema 3.59: *a*) sin correcciones y *b*) con correcciones de Sheppard (ver problema 5.30).

5.40 Los cuartos momentos respecto de la media de las dos distribuciones del problema 5.34 son 230 y 780, respectivamente. ¿Qué distribución se aproxima más a una distribución normal desde el punto de vista: *a*) de aplastamiento y *b*) del sesgo?

5.41 ¿Cuál de las distribuciones del problema 5.40 es: *a*) leptocúrtica, *b*) mesocúrtica y *c*) platicúrtica?

5.42 La desviación estándar de una distribución simétrica es 5. ¿Cuál deberá ser el valor del cuarto momento respecto de la media para que la distribución sea: *a*) leptocúrtica, *b*) mesocúrtica y *c*) platicúrtica?

5.43 *a*) Calcular el coeficiente percentil de curtosis, κ, de la distribución del problema 3.59.
b) Comparar su resultado con el valor teórico 0.263 para una distribución normal, e interpretar.
c) ¿Cómo se puede reconciliar este resultado con el del problema 5.39?

CÁLCULO DEL SESGO Y DE LA CURTOSIS EMPLEANDO SOFTWARE

5.44 Los datos de la figura 5-5 muestran un pronunciado pico en 50. Esto debe mostrarse en la medida de la curtosis de los datos. Empleando EXCEL, mostrar que el sesgo es prácticamente cero y que la curtosis es 2.0134.

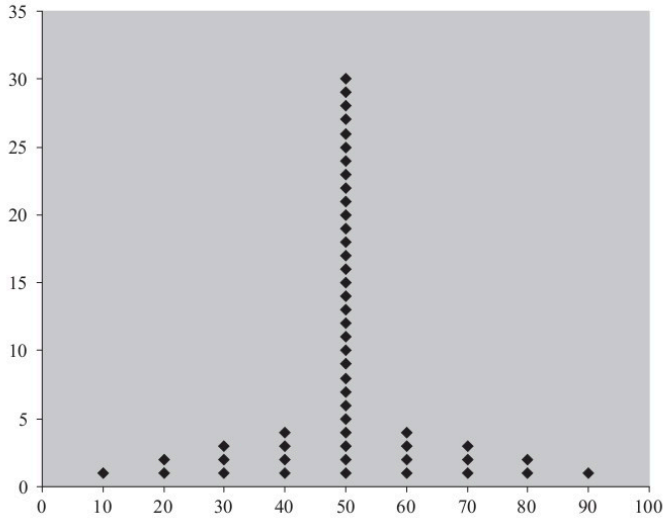


Figura 5-5 EXCEL, gráfica de los datos de puntuaciones de examen.

TEORÍA ELEMENTAL DE LA PROBABILIDAD

6

DEFINICIONES DE PROBABILIDAD

Definición clásica

Suponga que un evento E puede ocurrir en h de n maneras igualmente posibles. Entonces la probabilidad de que ocurra el evento (a la que se le llama *éxito*) se denota como

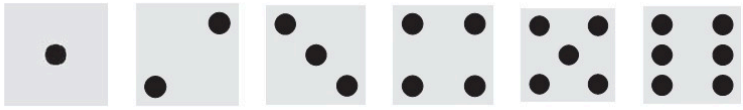
$$p = \Pr\{E\} = \frac{h}{n}$$

La probabilidad de que no ocurra el evento (a la que se le llama *fracaso*) se denota como

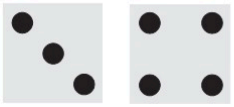
$$q = \Pr\{\text{no } E\} = \frac{n-h}{n} = 1 - \frac{h}{n} = 1 - p = 1 - \Pr\{E\}$$

Por lo tanto, $p + q = 1$ o bien $\Pr\{E\} + \Pr\{\text{no } E\} = 1$. El evento "no E" suele denotarse $\bar{E}$, $\tilde{E}$ o bien $\sim E$.

EJEMPLO 1 Cuando se lanza un dado, éste puede caer de seis maneras distintas.



Un evento E de que caiga un 3 o un 4 es:



y la probabilidad de E es $\Pr\{E\} = 2/6$ o bien $1/3$. La probabilidad de no obtener un 3 o un 4 (es decir, la probabilidad de obtener 1, 2, 5 o bien 6) es $\Pr\{\bar{E}\} = 1 - \Pr\{E\} = 2/3$.

Obsérvese que la probabilidad de un evento es un número entre 0 y 1. Si el evento no puede ocurrir, su probabilidad es 0. En cambio, si se trata de un evento que tiene que ocurrir (es decir, que es *seguro* que ocurra), su probabilidad es 1.

Si p es la probabilidad de que ocurra un evento, las *posibilidades* u *oportunidades* a favor de su ocurrencia son $p : q$ (que se lee "p a q"); las posibilidades en contra de que ocurra son $q : p$. Por lo tanto, las posibilidades en contra de que en un solo lanzamiento de un dado caiga un 3 o un 4 son $q : p = \frac{2}{3} : \frac{1}{3} = 2 : 1$ (es decir, 2 a 1).

Definición de frecuencia relativa

La definición clásica de probabilidad tiene la desventaja de que la expresión "igualmente posible" es vaga. Es más, como esta expresión parece ser sinónimo de "igualmente probable", la definición es *circular,* ya que está definiendo probabilidad en términos de probabilidad. Debido a esto, algunas personas han abogado por una definición estadística de probabilidad. De acuerdo con esto, se considera que la probabilidad estimada o *probabilidad empírica* de un evento es la *frecuencia relativa* de ocurrencia del evento cuando la cantidad de observaciones es muy grande. La probabilidad misma es el *límite* de esta frecuencia relativa a medida que la cantidad de observaciones aumenta de manera indefinida.

EJEMPLO 2 Si en 1 000 lanzamientos de una moneda se obtienen 529 caras, la frecuencia relativa con la que se obtienen caras es $529/1\,000 = 0.529$. Si en otros 1 000 lanzamientos se obtienen 493 caras, la frecuencia relativa en los 2 000 lanzamientos es $(529 + 493)/2\,000 = 0.511$. De acuerdo con la definición estadística, cada vez se estaría más cerca de un número que representa la probabilidad de que caiga cara en un lanzamiento de una sola moneda. Según los resultados presentados, este número sería 0.5 a una cifra significativa. Para obtener más cifras significativas se necesitan más observaciones.

La definición estadística, aunque útil en la práctica, tiene dificultades desde el punto de vista matemático, ya que puede ser que no exista un verdadero número límite. Debido a esto, la teoría de probabilidad moderna ha sido desarrollada en forma axiomática; es decir, el concepto de probabilidad se deja sin definir, que es lo mismo que ocurre en la geometría con los conceptos de *punto* y *línea*, que también se dejan sin definir.

PROBABILIDAD CONDICIONAL; EVENTOS INDEPENDIENTES Y DEPENDIENTES

Si E_1 y E_2 son dos eventos, la probabilidad de que ocurra E_2, dado que E_1 ha ocurrido, se denota $\Pr\{E_2|E_1\}$ o $\Pr\{E_2$ dado $E_1\}$ y se conoce como la *probabilidad condicional de* E_2 dado que E_1 ha ocurrido.

Si la ocurrencia o no ocurrencia de E_1 no afecta la probabilidad de ocurrencia de E_2, entonces $\Pr\{E_2|E_1\} = \Pr\{E_2\}$ y se dice que E_1 y E_2 son *eventos independientes*, de lo contrario se dice que son *eventos dependientes.*

Si se denota con E_1E_2 el evento de que "tanto E_1 como E_2 ocurran", evento al que suele llamarse *evento compuesto,* entonces

$$\Pr\{E_1E_2\} = \Pr\{E_1\}\Pr\{E_2|E_1\} \tag{1}$$

En particular,

$$\Pr\{E_1E_2\} = \Pr\{E_1\}\ \Pr\{E_2\} \qquad \text{para eventos independientes} \tag{2}$$

Para tres eventos E_1, E_2 y E_3, tenemos

$$\Pr\{E_1E_2E_3\} = \Pr\{E_1\}\ \Pr\{E_2|E_1\}\ \Pr\{E_3|E_1E_2\} \tag{3}$$

Es decir, la probabilidad de que ocurra E_1, E_2 y E_3 es igual a (la probabilidad de E_1) $\times$ (la probabilidad de E_2 dado que E_1 ha ocurrido) $\times$ (la probabilidad de E_3 dado que E_1 y E_2 han ocurrido). En particular,

$$\Pr\{E_1E_2E_3\} = \Pr\{E_1\}\ \Pr\{E_2\}\ \Pr\{E_3\} \qquad \text{para eventos independientes} \tag{4}$$

En general, si $E_1, E_2, E_3, \ldots, E_n$ son n eventos independientes que tienen probabilidades $p_1, p_2, p_3, \ldots, p_n$, entonces la probabilidad de que ocurra E_1 y E_2 y E_3 y $\cdots$ E_n es $p_1p_2p_3 \cdots p_n$.

EJEMPLO 3 Sean E_1 y E_2 los eventos "cae cara en el quinto lanzamiento" y "cae cara en el sexto lanzamiento" de una moneda, respectivamente. Entonces E_1 y E_2 son eventos independientes y, por lo tanto, la probabilidad de cara tanto en el quinto como en el sexto lanzamientos es (suponiendo que sea una moneda legal)

$$\Pr\{E_1E_2\} = \Pr\{E_1\}\ \Pr\{E_2\} = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4}$$

EJEMPLO 4 Si la probabilidad de que A esté vivo en 20 años es 0.7 y la probabilidad de que B esté vivo en 20 años es 0.5, entonces la probabilidad de que ambos estén vivos en 20 años es $(0.7)(0.5) = 0.35$.

EJEMPLO 5 Supóngase que una caja contiene 3 pelotas blancas y 2 pelotas negras. Sea E_1 el evento "la primera pelota que se saca es negra" y E_2 el evento "la segunda pelota que se saca es negra", donde las pelotas no se vuelvan a colocar en la caja una vez sacadas. Aquí E_1 y E_2 son eventos dependientes.

La probabilidad de que la primera pelota que se extraiga sea negra es $\Pr\{E_1\} = 2/(3+2) = \frac{2}{5}$. La probabilidad de que la segunda pelota que se extraiga sea negra, dado que la primera pelota que se extrajo fue negra, es $\Pr\{E_2|E_1\} = 1/(3+1) = \frac{1}{4}$. Por lo tanto, la probabilidad de que las dos pelotas que se extraigan sean negras es

$$\Pr\{E_1 E_2\} = \Pr\{E_1\}\Pr\{E_2|E_1\} = \frac{2}{5}\cdot\frac{1}{4} = \frac{1}{10}$$

EVENTOS MUTUAMENTE EXCLUYENTES

Se dice que dos o más eventos son *mutuamente excluyentes* si la ocurrencia de uno cualquiera de ellos excluye la ocurrencia de los otros. Entonces, si E_1 y E_2 son eventos mutuamente excluyentes, $\Pr\{E_1 E_2\} = 0$.

Si $E_1 + E_2$ denotan el evento "ocurre E_1 o E_2 o ambos", entonces

$$\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1 E_2\} \tag{5}$$

En particular,

$$\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\} \qquad \text{si los eventos son mutuamente excluyentes} \tag{6}$$

Por extensión se tiene que si $E_1, E_2, \ldots, E_n$ son n eventos mutuamente excluyentes que tienen probabilidades $p_1, p_2, \ldots, p_n$, entonces la probabilidad de que ocurran E_1 o E_2 o $\cdots E_n$ es $p_1 + p_2 + \cdots + p_n$.

La fórmula (*5*) también puede generalizarse a tres o más eventos mutuamente excluyentes.

EJEMPLO 6 Si E_1 es el evento "de una baraja se extrae un as" y E_2 es el evento "de una baraja se extrae un rey", entonces $\Pr\{E_1\} = \frac{4}{52} = \frac{1}{13}$ y $\Pr\{E_2\} = \frac{4}{52} = \frac{1}{13}$, y la probabilidad de en una sola extracción se extrae un as o un rey es

$$\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\} = \frac{1}{13} + \frac{1}{13} = \frac{2}{13}$$

ya que en una sola extracción o se extrae un as o se extrae un rey, y por lo tanto estos eventos son mutuamente excluyentes (figura 6-1).

A♣	2♣	3♣	4♣	5♣	6♣	7♣	8♣	9♣	10♣	J♣	Q♣	K♣
A♦	2♦	3♦	4♦	5♦	6♦	7♦	8♦	9♦	10♦	J♦	Q♦	K♦
A♥	2♥	3♥	4♥	5♥	6♥	7♥	8♥	9♥	10♥	J♥	Q♥	K♥
A♠	2♠	3♠	4♠	5♠	6♠	7♠	8♠	9♠	10♠	J♠	Q♠	K♠

Figura 6-1 E_1 es el evento "extraer un as" y E_2 es el evento "extraer un rey".

Obsérvese que E_1 y E_2 no tienen resultados en común. Estos eventos son mutuamente excluyentes.

EJEMPLO 7 Si E_1 es el evento "extraer un as" y E_2 es el evento "extraer una espada" de una baraja, E_1 y E_2 no son mutuamente excluyentes, pues se puede extraer el as de espadas (figura 6-2). Por lo tanto, la probabilidad de extraer un as o una espada o ambos es

$$\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1 E_2\} = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$

A♣	2♣	3♣	4♣	5♣	6♣	7♣	8♣	9♣	10♣	J♣	Q♣	K♣
A♦	2♦	3♦	4♦	5♦	6♦	7♦	8♦	9♦	10♦	J♦	Q♦	K♦
A♥	2♥	3♥	4♥	5♥	6♥	7♥	8♥	9♥	10♥	J♥	Q♥	K♥
A♠	2♠	3♠	4♠	5♠	6♠	7♠	8♠	9♠	10♠	J♠	Q♠	K♠

Figura 6-2 E_1 es el evento "extraer un as" y E_2 es el evento "extraer una espada".

Obsérvese que el evento "E_1 y E_2", que consta de los resultados en los que se den los dos eventos, es el as de espadas.

DISTRIBUCIONES DE PROBABILIDAD

Discretas

Si una variable X toma un conjunto discreto de valores $X_1, X_2, \ldots, X_K$ con probabilidades respectivas $p_1, p_2, \ldots, p_K$, donde $p_1 + p_2 + \cdots + p_K = 1$, esto se define como una *distribución de probabilidad discreta* de X. La función $p(X)$, que tiene los valores $p_1, p_2, \ldots, p_K$ para $X = X_1, X_2, \ldots, X_K$, respectivamente, se llama *función de probabilidad* o *función de frecuencia* de X. Como X puede tomar ciertos valores con determinadas probabilidades, suele llamársele *variable aleatoria discreta*. A las variables aleatorias también se les conoce como *variables estocásticas*.

EJEMPLO 8 Se lanza un par de dados; sea X la suma de los puntos obtenidos en estos dos dados. La distribución de probabilidad es la que se muestra en la tabla 6.1. Por ejemplo, la probabilidad de que la suma sea 5 es $\frac{4}{36} = \frac{1}{9}$; así que de 900 veces que se lancen los dos dados se espera que en 100 la suma de los puntos sea 5.

Obsérvese la analogía con las distribuciones de frecuencias relativas empleando probabilidades en lugar de frecuencias relativas. De manera que las distribuciones de probabilidad pueden considerarse como formas teóricas o formas límites ideales de las distribuciones de frecuencias relativas cuando la cantidad de observaciones es muy grande. A esto se debe que las distribuciones de probabilidad se consideren distribuciones de *poblaciones*, mientras que las distribuciones de frecuencias relativas son distribuciones de *muestras* obtenidas de estas poblaciones.

Tabla 6.1

X	2	3	4	5	6	7	8	9	10	11	12
$p(X)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Una distribución de probabilidad se puede representar graficando $p(X)$ contra X, como se hace con las distribuciones de frecuencias relativas (ver problema 6.11).

Con probabilidades acumuladas se obtienen *distribuciones de probabilidad acumulada*, que son análogas a las distribuciones de frecuencia relativa acumulada. A las funciones correspondientes a estas distribuciones se les suele llamar *funciones de distribución*.

La distribución de la tabla 6.1 puede obtenerse empleando EXCEL. La porción de una hoja de cálculo de EXCEL que se muestra a continuación se obtiene ingresando Dado 1 en A1, Dado 2 en B1 y Suma en C1. Los 36 resultados posibles al lanzar dos dados se ingresan en A2:B37. En C2 se ingresa =SUM(A2:B2), se da clic y se arrastra desde C2 hasta C37. Observando que la suma 2 se obtiene una vez; la suma 3, dos veces, etc., se forma la distribución de probabilidad de la tabla 6.1.

Dado 1	Dado 2	Suma
1	1	2
1	2	3
1	3	4
1	4	5
1	5	6
1	6	7
2	1	3
2	2	4
2	3	5
2	4	6
2	5	7

Dado 1	Dado 2	Suma
2	6	8
3	1	4
3	2	5
3	3	6
3	4	7
3	5	8
3	6	9
4	1	5
4	2	6
4	3	7
4	4	8
4	5	9
4	6	10
5	1	6
5	2	7
5	3	8
5	4	9
5	5	10
5	6	11
6	1	7
6	2	8
6	3	9
6	4	10
6	5	11
6	6	12

Continua

Las ideas anteriores pueden extenderse al caso en el que la variable X puede tomar un conjunto continuo de valores. El polígono de frecuencias relativas de la muestra se convierte, en el caso teórico o límite de una población, en una curva continua (como la que se muestra en la figura 6-3) cuya ecuación es $Y = p(X)$. El área total limitada por el eje X, bajo esta curva, es igual a 1, y el área entre las rectas $X = a$ y $X = b$ (que aparece sombreada en la figura 6-3) corresponde a la probabilidad de que X se encuentre entre a y b, lo que se denota como $\Pr\{a < X < b\}$.

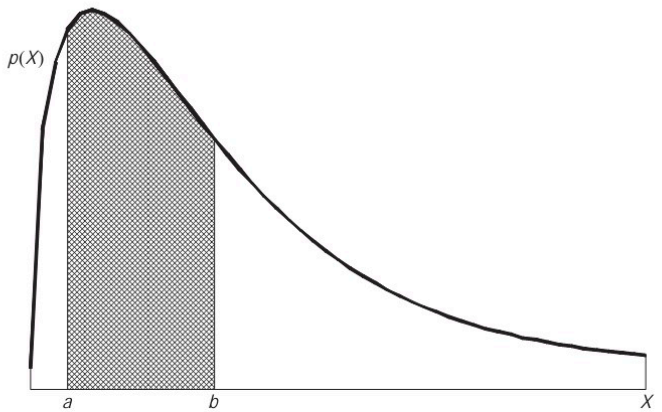


Figura 6-3 $\Pr\{a < X < b\}$ es el área sombreada bajo la función de densidad.

A $p(X)$ se le conoce como *función de densidad de probabilidad* o brevemente *función de densidad*, y cuando se da una de estas funciones se dice que se define una *distribución de probabilidad continua* para X; a la variable X suele llamársele *variable aleatoria continua*.

Como en el caso discreto, se pueden definir distribuciones de probabilidad acumulada y las funciones de distribución correspondientes.

ESPERANZA MATEMÁTICA

Si p es la probabilidad de que una persona reciba una cantidad de dinero S, pS define la *esperanza matemática* (o simplemente la *esperanza*).

EJEMPLO 9 Encontrar $E(X)$ para la distribución de la suma de los dos dados dada en la tabla 6.1. La distribución se presenta en los siguientes resultados de EXCEL. En A2:B12 se presenta la distribución en la que los valores $p(X)$ se han convertido a la forma decimal. En C2 se ingresa la expresión =A2*B2, se da clic y se arrastra desde C2 hasta C12. En C13 se ingresa la expresión =Sum(C2:C12) y se obtiene la esperanza matemática, que es 7.

X	$p(X)$	$XP(X)$
2	0.027778	0.055556
3	0.055556	0.166667
4	0.083333	0.333333
5	0.111111	0.555556
6	0.138889	0.833333
7	0.166667	1.166667
8	0.138889	1.111111
9	0.111111	1
10	0.083333	0.833333
11	0.055556	0.611111
12	0.027778	0.333333
		7

El concepto de esperanza es fácil de extender. Si X denota una variable aleatoria discreta que puede tomar los valores $X_1, X_2, \ldots, X_K$ con probabilidades $p_1, p_2, \ldots, p_K$, respectivamente, donde $p_1 + p_2 + \cdots + p_K = 1$, la *esperanza matemática* de X (o simplemente la *esperanza* de X), que se denota $E(X)$, se define de la manera siguiente:

$$E(X) = p_1X_1 + p_2X_2 + \cdots + p_KX_K = \sum_{j=1}^{K} p_jX_j = \sum pX \tag{7}$$

Si en esta esperanza se sustituyen las probabilidades p_j por las frecuencias relativas f_j/N, donde $N = \sum f_j$, la esperanza se reduce a $(\sum fX)/N$, que es la media aritmética $\bar{X}$ de una muestra de tamaño N en la que $X_1, X_2, \ldots, X_K$ se presentan con estas frecuencias relativas. A medida que N se vuelve cada vez más grande, las frecuencias relativas f_j/N se aproximan a las probabilidades p_j. Esto lleva a interpretar $E(X)$ como la media de la población de la que ha sido tomada la muestra. Si se denota con m a la media muestral, a la media poblacional se le denota con la correspondiente letra griega, μ (mu).

La esperanza también puede ser definida para variables aleatorias continuas, pero esta definición requiere el uso del cálculo.

RELACIÓN ENTRE MEDIA Y VARIANZA POBLACIONALES Y MUESTRALES

Si de una población se toma en forma aleatoria una muestra de tamaño N (es decir, de manera que todas las muestras de tamaño N sean igualmente probables), se puede demostrar que el *valor esperado para la media muestral m es la media poblacional μ*.

Sin embargo, no se sigue que el valor esperado de cualquier cantidad calculada a partir de una muestra sea la cantidad poblacional correspondiente. Por ejemplo, el valor esperado de la varianza muestral, como se ha definido aquí, no es la varianza poblacional, sino $(N-1)/N$ veces esta varianza. A esto se debe que algunos especialistas en estadística prefieran definir la varianza muestral como la varianza aquí definida pero multiplicada por $N/(N-1)$.

ANÁLISIS COMBINATORIO

Para obtener probabilidades de eventos complejos, hacer una enumeración de los casos suele ser difícil, tedioso o ambas cosas. Para facilitar esta tarea se hace uso de los principios básicos de una materia llamada *análisis combinatorio.*

Principio fundamental

Si un evento puede ocurrir de n_1 maneras diferentes, y si una vez que ha ocurrido, otro evento puede ocurrir de n_2 maneras diferentes, entonces la cantidad de maneras en que pueden ocurrir los dos eventos, en este orden específico, es $n_1 n_2$.

EJEMPLO 10 En una hoja de cálculo de EXCEL se ingresan los números 0 a 5 en las casillas A1 a A6. En B1 se ingresa =FACT(A1), se hace clic y se arrastra desde B1 hasta B6. Después, para graficar los puntos, se emplea el asistente para gráficos de EXCEL. La función =FACT(n) es lo mismo que $n!$ Para $n = 0, 1, 2, 3, 4$ y 5 =FACT(n) es igual a 1, 1, 2, 6, 24 y 120. La figura 6-4 se generó con el asistente para gráficos de EXCEL.

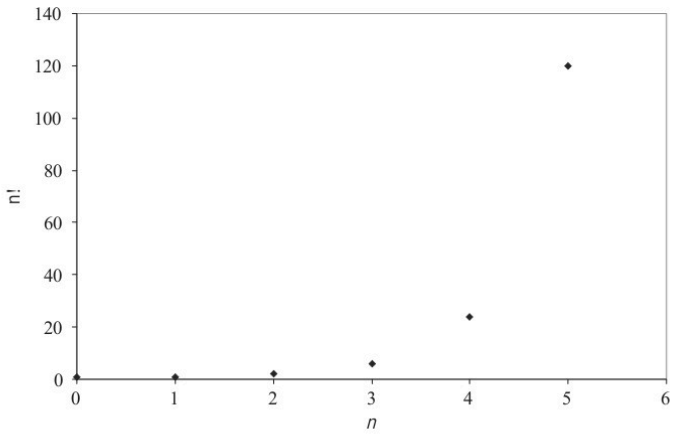


Figura 6-4 Gráfica de *n*! generada con EXCEL.

EJEMPLO 11 La cantidad de permutaciones de las letras *a*, *b* y *c* tomadas de dos en dos es ${}_3P_2 = 3 \cdot 2 = 6$. Estas permutaciones son *ab*, *ba*, *ac*, *ca*, *bc* y *cb*.

El número de permutaciones de n objetos de los cuales n_1 son iguales, n_2 son iguales, ... es

$$\frac{n!}{n_1!\, n_2! \cdots} \qquad \text{donde } n = n_1 + n_2 + \cdots \tag{10}$$

EJEMPLO 12 El número de permutaciones de las letras en la palabra *statistics* es

$$\frac{10!}{3!\,3!\,1!\,2!\,1!} = 50\,400$$

ya que hay 3 *eses*, 3 *tes*, 1 *a*, 2 *íes* y 1 *c*.

COMBINACIONES

Una combinación de n objetos diferentes tomados de r en r es una selección de r de los n objetos sin importar el orden. El número de combinaciones de n objetos tomados de r en r se denota mediante el símbolo $\binom{n}{r}$ y está dado por

$$\binom{n}{r} = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!} \tag{11}$$

EJEMPLO 13 El número de combinaciones que se pueden hacer con las letras *a*, *b*, y *c*, tomadas de dos en dos, es

$$\binom{3}{2} = \frac{3 \cdot 2}{2!} = 3$$

Estas combinaciones son *ab*, *ac* y *bc*. Obsérvese que *ab* es la misma combinación que *ba*, pero no la misma permutación.

Con EXCEL, las combinaciones de 3 objetos tomando 2 a la vez se obtienen con el comando =COMBIN(3,2), que da como resultado 3.

APROXIMACIÓN DE STIRLING PARA *n*!

Cuando n es grande es poco práctico evaluar directamente $n!$ En tales casos se hace uso de una fórmula de aproximación desarrollada por James Stirling:

$$n! \approx \sqrt{2\pi n}\, n^n e^{-n} \tag{12}$$

donde $e = 2.71828\ldots$ es la base del logaritmo natural (ver problema 6.31).

RELACIÓN ENTRE LA PROBABILIDAD Y LA TEORÍA DE CONJUNTOS

Como se ve en la figura 6-5, un *diagrama de Venn* representa, mediante un rectángulo, todos los resultados posibles de un *experimento*, a lo que se llama el *espacio muestral S*. Los eventos se representan como figuras tetraédricas o como círculos dentro del espacio muestral. Si S contiene únicamente una cantidad finita de puntos, entonces a cada punto se le puede asociar un número no negativo, llamado *probabilidad*, de manera que la suma de todos los números correspondientes a los puntos de S sea 1. Un evento es un conjunto (o una colección) de puntos en S, como los indicados en la figura 6-5 por E_1 y E_2.

DIAGRAMAS DE EULER O DE VENN Y PROBABILIDAD

El evento $E_1 + E_2$ es el conjunto de puntos que están *en* E_1 *o en* E_2 *o en ambos*, mientras que el evento E_1E_2 es el conjunto de puntos que son *comunes a ambos,* E_1 *y* E_2. La probabilidad de un evento por ejemplo el evento E_1 es la suma de las probabilidades correspondientes a todos los puntos que están en el conjunto E_1. De igual manera, la probabilidad de $E_1 + E_2$, que se denota $\Pr\{E_1 + E_2\}$, es la suma de las probabilidades correspondientes a todos los puntos contenidos en el conjunto $E_1 + E_2$. Si E_1 y E_2 no tienen puntos en común (es decir, si los eventos son mutuamente excluyentes), entonces $\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\}$. Si tienen puntos en común, entonces $\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1E_2\}$.

El conjunto $E_1 + E_2$ también suele denotarse $E_1 \cup E_2$ y se conoce como la *unión* de los dos conjuntos. El conjunto E_1E_2 también suele denotarse $E_1 \cap E_2$ y se conoce como la *intersección* de los dos conjuntos. Se pueden hacer extensiones a más de dos conjuntos; así, en lugar de $E_1 + E_2 + E_3$ y $E_1E_2E_3$, se pueden emplear las notaciones $E_1 \cup E_2 \cup E_3$ y $E_1 \cap E_2 \cap E_3$, respectivamente.

DIAGRAMAS DE EULER O DE VENN Y PROBABILIDAD

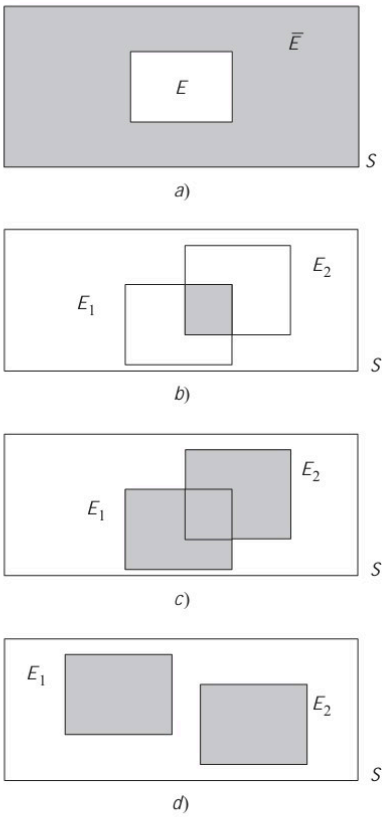


Figura 6-5 Operaciones con eventos. *a*) El complemento del evento E aparece sombreado y se denota $\bar{E}$; *b*) la intersección de los eventos E_1 y E_2 aparece sombreada y se escribe $E_1 \cap E_2$; *c*) la unión de los eventos E_1 y E_2 aparece sombreada y se denota $E_1 \cup E_2$; *d*) los eventos E_1 y E_2 son mutuamente excluyentes, es decir, $E_1 \cap E_2 = \phi$.

El símbolo ϕ (la letra *fi* del alfabeto griego) suele emplearse para denotar el conjunto que no tiene ningún punto, conjunto al que se le conoce como *conjunto vacío*. La probabilidad que se le asigna a un evento que corresponde a este conjunto es cero (es decir, $\Pr\{\phi\} = 0$). Si E_1 y E_2 no tienen puntos en común, se escribe $E_1E_2 = \phi$, lo que significa que los eventos correspondientes son mutuamente excluyentes, por lo que $\Pr\{E_1E_2\} = 0$.

Con esta visión moderna, una variable aleatoria es una función definida en cada punto de un espacio muestral. Por ejemplo, en el problema 6.37, la variable aleatoria es la suma de las coordenadas de cada punto.

Empleando conceptos del cálculo pueden extenderse las ideas anteriores al caso en el que S tenga una cantidad infinita de puntos.

EJEMPLO 14 Un experimento consiste en lanzar un par de dados. El evento E_1 es que se obtenga un 7, es decir, que la suma de los puntos en los dados sea 7. El evento E_2 es que en el dado 1 se obtenga un número non. A continuación se presenta el espacio muestral S y los eventos E_1 y E_2. Encontrar $\Pr\{E_1\}$, $\Pr\{E_2\}$, $\Pr\{E_1 \cap E_2\}$ y $\Pr\{E_1 \cup E_2\}$. En la figura 6.6 se presentan los resultados de MINITAB con E_1, E_2 y S en rectángulos separados.

$$P(E_1) = 6/36 = 1/6 \quad P(E_2) = 18/36 = 1/2 \quad P(E_1 \cap E_2) = 3/36 = 1/12$$
$$P(E_1 \cup E_2) = 6/36 + 18/36 - 3/36 = 21/36.$$

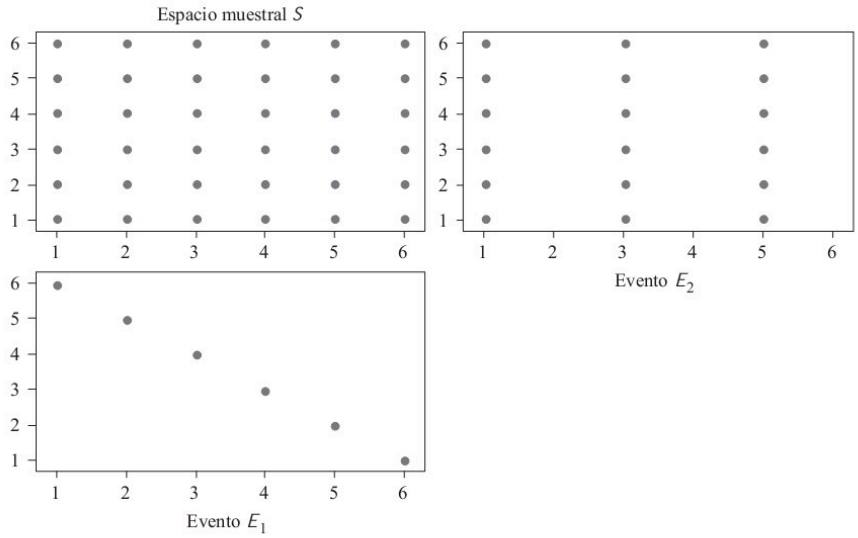


Figura 6-6 Resultados de MINITAB para el ejemplo 14.

PROBLEMAS RESUELTOS

REGLAS FUNDAMENTALES DE LA PROBABILIDAD

6.1 Determinar o estimar la probabilidad p de cada uno de los eventos siguientes:

a) Al lanzar una vez un dado obtener un número non.
b) Al lanzar dos veces una moneda obtener por lo menos una cara.
c) Al sacar una carta de una baraja, bien barajada, con 52 cartas obtener un as, un 10 de diamantes o un 2 de espadas.
d) Al lanzar una vez un par de dados su suma sea siete.
e) Si en 100 lanzamientos de una moneda se obtuvieron 56 caras, en el siguiente lanzamiento obtener una cruz.

SOLUCIÓN

a) De seis casos equiprobables posibles, tres casos (que caiga 1, 3 o 5) son favorables al evento. Por lo tanto, $p = \frac{3}{6} = \frac{1}{2}$.

b) Si H denota "cara" y T denota "cruz", en los dos lanzamientos se pueden obtener los casos siguientes: HH, HT, TH, TT, todos igualmente posibles. De éstos, sólo los tres primeros son favorables al evento. Por lo tanto, $p = \frac{3}{4}$.

c) Este evento puede darse de seis maneras (as de espadas, as de corazones, as de tréboles, as de diamantes, 10 de diamantes y 2 de espadas) en los 52 casos igualmente posibles. Por lo tanto, $p = \frac{6}{52} = \frac{3}{26}$.

d) Cada una de las caras de un dado puede relacionarse con cada una de las seis caras del otro dado, de manera que la cantidad de casos que pueden presentarse, todos igualmente posibles, es $6 \cdot 6 = 36$. Estos casos se pueden denotar (1, 1), (2, 1), (3, 1), . . . , (6, 6).

Hay seis formas de obtener la suma de 7, denotada por (1, 6), (2, 5), (3, 4), (4, 3), (5, 2) y (6, 1). Por lo tanto, $p = \frac{6}{36} = \frac{1}{6}$.

e) Como en 100 lanzamientos se obtuvieron $100 - 56 = 44$ cruces, la *probabilidad estimada* (o *empírica*) de que caiga cruz es la frecuencia relativa $44/100 = 0.44$.

6.2 Un experimento consiste en lanzar una moneda y un dado. Si E_1 es el evento en que se obtenga "cara" al lanzar la moneda y E_2 es el evento en que se obtenga "3 o 6" al lanzar el dado, expresar en palabras cada uno de los eventos siguientes:

a) $\bar{E}_1$ *c*) $E_1 E_2$ *e*) $\Pr\{E_1 | E_2\}$
b) $\bar{E}_2$ *d*) $\Pr\{E_1 \bar{E}_2\}$ *f*) $\Pr\{\bar{E}_1 + \bar{E}_2\}$

SOLUCIÓN

a) Cruz en la moneda y cualquier cosa en el dado.
b) 1, 2, 4 ó 5 en el dado y cualquier cosa en la moneda.
c) Cara en la moneda y 3 ó 6 en el dado.
d) Probabilidad de cara en la moneda y 1, 2, 4 ó 5 en el dado.
e) Probabilidad de cara en la moneda, dado que en el dado se obtuvo 3 ó 6.
f) Probabilidad de cruz en la moneda o 1, 2, 4 ó 5 en el dado o ambas cosas

6.3 De una caja que contiene 6 pelotas rojas, 4 pelotas blancas y 5 pelotas azules se extrae, de manera aleatoria, una pelota. Determinar la probabilidad de que la pelota extraída sea: *a*) roja, *b*) blanca, *c*) azul, *d*) no sea roja y *e*) sea roja o blanca.

SOLUCIÓN

*C*on *R*, *W* y *B* se denotan los eventos de que la pelota que se saque sea roja, blanca o azul, respectivamente. Entonces:

a) $$\Pr\{R\} = \frac{\text{maneras de sacar una pelota roja}}{\text{total de maneras de sacar una pelota}} = \frac{6}{6+4+5} = \frac{6}{15} = \frac{2}{5}$$

b) $$\Pr\{W\} = \frac{4}{6+4+5} = \frac{4}{15}$$

c) $$\Pr\{B\} = \frac{5}{6+4+5} = \frac{5}{15} = \frac{1}{3}$$

d) $$\Pr\{\bar{R}\} = 1 - \Pr\{R\} = 1 - \frac{2}{5} = \frac{3}{5} \qquad \text{de acuerdo con el inciso } a)$$

e) $$\Pr\{R + W\} = \frac{\text{maneras de sacar una pelota roja o una pelota blanca}}{\text{total de maneras de sacar una pelota}} = \frac{6+4}{6+4+5} = \frac{10}{15} = \frac{2}{3}$$

Otro método

$$\Pr\{R + W\} = \Pr\{\bar{B}\} = 1 - \Pr\{B\} = 1 - \frac{1}{3} = \frac{2}{3} \qquad \text{de acuerdo con el inciso } c)$$

Obsérvese que $\Pr\{R + W\} = \Pr\{R\} + \Pr\{W\}$ (es decir, $\frac{2}{3} = \frac{2}{5} + \frac{4}{15}$). Éste es un ejemplo de la regla general $\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\}$ válida para eventos E_1 y E_2 *mutuamente excluyentes*.

6.4 Un dado se lanza dos veces. Encontrar la probabilidad de obtener un 4, un 5 o un 6 en el primer lanzamiento y un 1, 2, 3 ó 4 en el segundo lanzamiento.

SOLUCIÓN

Sea E_1 el evento "4, 5 ó 6" en el primer lanzamiento y E_2 el evento "1, 2, 3 ó 4" en el segundo lanzamiento. A cada una de las seis maneras en que puede caer el dado en el primer lanzamiento se le asocia cada una de las seis maneras en que puede caer en el segundo lanzamiento, lo que hace un total de $6 \cdot 6 = 36$ maneras, todas igualmente probables. A cada una de las tres manera en que puede ocurrir E_1 se le asocia cada una de las cuatro maneras en que puede ocurrir E_2, obteniéndose $3 \cdot 4 = 12$ maneras en las que pueden ocurrir E_1 y E_2 o $E_1 E_2$. Por lo tanto, $\Pr\{E_1 E_2\} = 12/36 = 1/3$.

Obsérvese que $\Pr\{E_1 E_2\} = \Pr\{E_1\}\Pr\{E_2\}$ (es decir, $\frac{1}{3} = \frac{3}{6} \cdot \frac{4}{6}$) es válido para *eventos independientes* E_1 y E_2.

6.5 De una baraja, bien barajada, con 52 cartas se extraen dos cartas. Encuentre la probabilidad de que las dos sean ases si la primera carta: *a*) se devuelve a la baraja y *b*) no se devuelve a la baraja.

SOLUCIÓN

Sea $E_1 =$ evento "as" en la primera extracción y sea $E_2 =$ evento "as" en la segunda extracción.

a) Si la primera carta se devuelve a la baraja, E_1 y E_2 son eventos independientes. Por lo tanto, Pr{las dos cartas extraídas sean ases} $= \Pr\{E_1E_2\} = \Pr\{E_1\}\ \Pr\{E_2\} = (\frac{4}{52})(\frac{4}{52}) = \frac{1}{169}$.

b) La primera carta se puede extraer de 52 maneras y la segunda puede extraerse de 51 maneras, ya que la primera carta no se devuelve a la baraja. Por lo tanto, las dos cartas se pueden extraer de $52 \cdot 51$ maneras todas igualmente posibles.

Hay cuatro maneras en las que puede ocurrir E_1 y tres maneras en las que E_2 puede ocurrir, de manera que E_1 y E_2 o E_1E_2 puede ocurrir de $4 \cdot 3$ maneras. Por lo tanto, $\Pr\{E_1E_2\} = (4 \cdot 3)/(52 \cdot 51) = \frac{1}{221}$.

Obsérvese que $\Pr\{E_2|E_1\} =$ Pr{segunda carta sea un as dado que la primera carta es un as} $= \frac{3}{51}$. De manera que este resultado ilustra la regla general $\Pr\{E_1E_2\} = \Pr\{E_1\}\ \Pr\{E_2|E_1\}$ donde E_1 y E_2 son eventos dependientes.

6.6 De la caja del problema 6.3 se extraen, sucesivamente, tres pelotas. Encuéntrese la probabilidad de que se extraigan en el orden roja, blanca y azul: *a*) si cada pelota se devuelve a la caja y *b*) si no se devuelve.

SOLUCIÓN

Sea $R =$ evento "roja" en la primera extracción, $W =$ evento "blanca" en la segunda extracción y $B =$ evento "azul" en la tercera extracción. Lo que se busca es Pr{*RWB*}.

a) Si cada una de las pelotas se devuelve, entonces *R*, *W* y *B* son eventos independientes y

$$\Pr\{RWB\} = \Pr\{R\}\ \Pr\{W\}\ \Pr\{B\} = \left(\frac{6}{6+4+5}\right)\left(\frac{4}{6+4+5}\right)\left(\frac{5}{6+4+5}\right) = \left(\frac{6}{15}\right)\left(\frac{4}{15}\right)\left(\frac{5}{15}\right) = \frac{8}{225}$$

b) Si las pelotas no se devuelven, entonces *R*, *W* y *B* son eventos dependientes y

$$\Pr\{RWB\} = \Pr\{R\}\ \Pr\{W|R\}\ \Pr\{B|WR\} = \left(\frac{6}{6+4+5}\right)\left(\frac{4}{5+4+5}\right)\left(\frac{5}{5+3+5}\right)$$
$$= \left(\frac{6}{15}\right)\left(\frac{4}{14}\right)\left(\frac{5}{13}\right) = \frac{4}{91}$$

donde Pr{*B*|*WR*} es la probabilidad condicional de extraer una pelota azul si se han extraído ya una roja y una blanca.

6.7 Encuéntrese la probabilidad de que en dos lanzamientos de un dado se obtenga por lo menos un 4.

SOLUCIÓN

Sea $E_1 =$ el evento "4" en el primer lanzamiento, $E_2 =$ el evento "4" en el segundo lanzamiento y $E_1 + E_2 =$ el evento "4" en el primer lanzamiento o "4" en el segundo lanzamiento o ambos = el evento de obtener por lo menos un 4. Lo que se busca es $\Pr\{E_1 + E_2\}$.

Primer método

Los dos dados pueden caer en un total de $6 \cdot 6 = 36$ maneras igualmente posibles. Además,

Cantidad de maneras en las que puede ocurrir E_1 pero no $E_2 = 5$
Cantidad de maneras en las que puede ocurrir E_2 pero no $E_1 = 5$
Cantidad de maneras en las que pueden ocurrir E_1 y $E_2 = 1$

Por lo tanto, la cantidad de maneras en las que puede ocurrir por lo menos uno de los eventos E_1 o E_2 es $5 + 5 + 1 = 11$, y por lo tanto, $\Pr\{E_1 + E_2\} = \frac{11}{36}$.

Segundo método

Como E_1 y E_2 no son mutuamente excluyentes, $\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1E_2\}$. Además, como E_1 y E_2 son independientes, $\Pr\{E_1E_2\} = \Pr\{E_1\}\ \Pr\{E_2\}$. Por lo tanto,

$$\Pr\{E_1 + E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1\}\ \Pr\{E_2\} = \tfrac{1}{6} + \tfrac{1}{6} - (\tfrac{1}{6})(\tfrac{1}{6}) = \tfrac{11}{36}.$$

Tercer método

$$\Pr\{\text{obtener por lo menos un 4}\} + \Pr\{\text{no obtener ningún 4}\} = 1.$$

Por lo tanto,
$$\begin{aligned}\Pr\{\text{obtener por lo menos un 4}\} &= 1 - \Pr\{\text{no obtener ningún 4}\}\\ &= 1 - \Pr\{\text{no obtener un 4 ni en el primero ni en el segundo lanzamiento}\}\\ &= 1 - \Pr\{\bar{E}_1\bar{E}_2\} = 1 - \Pr\{\bar{E}_1\}\ \Pr\{\bar{E}_2\}\\ &= 1 - (\tfrac{5}{6})(\tfrac{5}{6}) = \tfrac{11}{36}\end{aligned}$$

6.8 Una bolsa contiene 4 pelotas blancas y 2 pelotas negras; otra contiene 3 pelotas blancas y 5 pelotas negras. Si se saca una pelota de cada bolsa, encontrar la probabilidad de que: *a*) ambas sean blancas, *b*) ambas sean negras y *c*) una sea blanca y la otra sea negra.

SOLUCIÓN

Sea W_1 = el evento "blanca" de la primera bolsa y W_2 = el evento "blanca" de la segunda bolsa.

a) $$\Pr\{W_1W_2\} = \Pr\{W_1\}\ \Pr\{W_2\} = \left(\frac{4}{4+2}\right)\left(\frac{3}{3+5}\right) = \frac{1}{4}$$

b) $$\Pr\{\bar{W}_1\bar{W}_2\} = \Pr\{\bar{W}_1\}\ \Pr\{\bar{W}_2\} = \left(\frac{2}{4+2}\right)\left(\frac{5}{3+5}\right) = \frac{5}{24}$$

c) El evento "una es blanca y la otra es negra" es lo mismo que el evento "o la primera es blanca y la segunda es negra *o* la primera es negra y la segunda es blanca"; es decir, $W_1\bar{W}_2 + \bar{W}_1W_2$. Como los eventos $W_1\bar{W}_2$ y $\bar{W}_1W_2$ son mutuamente excluyentes, se tiene

$$\begin{aligned}\Pr\{W_1\bar{W}_2 + \bar{W}_1W_2\} &= \Pr\{W_1\bar{W}_2\} + \Pr\{\bar{W}_1W_2\}\\ &= \Pr\{W_1\}\ \Pr\{\bar{W}_2\} + \Pr\{\bar{W}_1\}\ \Pr\{W_2\}\\ &= \left(\frac{4}{4+2}\right)\left(\frac{5}{3+5}\right) + \left(\frac{2}{4+2}\right)\left(\frac{3}{3+5}\right) = \frac{13}{24}\end{aligned}$$

Otro método

La probabilidad que se busca es $1 - \Pr\{W_1W_2\} - \Pr\{\bar{W}_1\bar{W}_2\} = 1 - \frac{1}{4} - \frac{5}{24} = \frac{13}{24}$.

6.9 *A* y *B* juegan 12 partidos de ajedrez, de los cuales, *A* gana 6, *B* gana 4 y en 2 terminan empatados. Se ponen de acuerdo para jugar otros 3 partidos. Encuéntrese la probabilidad de que: *a*) *A* gane los tres partidos, *b*) 2 partidos terminen empatados, *c*) *A* y *B* ganen alternadamente y *d*) *B* gane por lo menos un partido.

SOLUCIÓN

Sean A_1, A_2 y A_3 los eventos "*A* gana" en el primero, el segundo y el tercer partidos, respectivamente; sean B_1, B_2 y B_3 los eventos "*B* gana" el primero, el segundo y el tercer partidos, respectivamente, y D_1, D_2 y D_3 los eventos "terminan empatados" en el primero, el segundo y el tercer partidos, respectivamente.

Sobre la base de experiencias anteriores (probabilidad empírica), asumir que $\Pr\{A \text{ gana cualquiera de los partidos}\} = \frac{6}{12} = \frac{1}{2}$, que $\Pr\{B \text{ gana cualquiera de los partidos}\} = \frac{4}{12} = \frac{1}{3}$ y que $\Pr\{\text{termina en empate en cualquier partido}\} = \frac{2}{12} = \frac{1}{6}$.

a) $\Pr\{A \text{ gane los tres partidos}\} = \Pr\{A_1A_2A_3\} = \Pr\{A_1\}\ \Pr\{A_2\}\ \Pr\{A_3\} = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{8}$

suponiendo que el resultado de cada partido sea independiente de los resultados de los otros partidos, lo que parece razonable (a menos, por supuesto, que el jugador se vea *psicológicamente influenciado* por los otros partidos ganados o perdidos).

b) $\Pr\{\text{2 partidos terminen empatados}\} = \Pr\{\text{el primero y el segundo } o \text{ el primero y el tercero } o \text{ el segundo y el tercer partidos terminen empatados}\}$

$$= \Pr\{D_1 D_2 \bar{D}_3\} + \Pr\{D_1 \bar{D}_2 D_3\} + \Pr\{\bar{D}_1 D_2 D_3\}$$
$$= \Pr\{D_1\}\ \Pr\{D_2\} \Pr\{\bar{D}_3\} + \Pr\{D_1\} \Pr\{\bar{D}_2\} \Pr\{D_3\}$$
$$+ \Pr\{\bar{D}_1\} \Pr\{D_2\}\ \Pr\{D_3\}$$
$$= \left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\left(\frac{5}{6}\right) + \left(\frac{1}{6}\right)\left(\frac{5}{6}\right)\left(\frac{1}{6}\right) + \left(\frac{5}{6}\right)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{15}{216} = \frac{5}{72}$$

c) $\Pr\{A \text{ y } B \text{ ganen alternadamente}\} = \Pr\{A \text{ gane y después } B \text{ gane y después } A \text{ gane } o \text{ que } B \text{ gane y después } A \text{ gane y después } B \text{ gane}\}$

$$= \Pr\{A_1 B_2 A_3 + B_1 A_2 B_3\} = \Pr\{A_1 B_2 A_3\} + \Pr\{B_1 A_2 B_3\}$$
$$= \Pr\{A_1\}\ \Pr\{B_2\}\ \Pr\{A_3\} + \Pr\{B_1\}\ \Pr\{A_2\}\ \Pr\{B_3\}$$
$$= \left(\frac{1}{2}\right)\left(\frac{1}{3}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{2}\right)\left(\frac{1}{3}\right) = \frac{5}{36}$$

d) $\Pr\{B \text{ gane por lo menos un partido}\} = 1 - \Pr\{B \text{ no gane ningún partido}\}$

$$= 1 - \Pr\{\bar{B}_1 \bar{B}_2 \bar{B}_3\} = 1 - \Pr\{\bar{B}_1\}\ \Pr\{\bar{B}_2\}\ \Pr\{\bar{B}_3\}$$
$$= 1 - \left(\frac{2}{3}\right)\left(\frac{2}{3}\right)\left(\frac{2}{3}\right) = \frac{19}{27}$$

DISTRIBUCIONES DE PROBABILIDAD

6.10 Encontrar la probabilidad de que haya niños o niñas en familias con tres hijos, suponiendo probabilidades iguales para niños que para niñas.

SOLUCIÓN

Sea $B =$ el evento "niño en la familia" y $G =$ el evento "niña en la familia". De acuerdo con la suposición de probabilidades iguales, $\Pr\{B\} = \Pr\{G\} = \frac{1}{2}$. En las familias con tres hijos pueden presentarse los siguientes eventos mutuamente excluyentes con las probabilidades que se indican.

a) Tres niños (*BBB*):

$$\Pr\{BBB\} = \Pr\{B\}\ \Pr\{B\}\ \Pr\{B\} = \frac{1}{8}$$

Aquí se supone que el que nazca un niño no está influenciado de manera alguna porque el hijo anterior haya sido también niño, es decir, se supone que los eventos son independientes.

b) Tres niñas (*GGG*): Como en el inciso *a*) o por simetría,

$$\Pr\{GGG\} = \frac{1}{8}$$

c) Dos niños y una niña (*BBG* + *BGB* + *GBB*):

$$\Pr\{BBG + BGB + GBB\} = \Pr\{BBG\} + \Pr\{BGB\} + \Pr\{GBB\}$$
$$= \Pr\{B\}\ \Pr\{B\}\ \Pr\{G\} + \Pr\{B\}\ \Pr\{G\}\ \Pr\{B\} + \Pr\{G\}\ \Pr\{B\}\ \Pr\{B\}$$
$$= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

d) Dos niñas y un niño (*GGB* + *GBG* + *BGG*): como en el inciso *c*) o por simetría, la probabilidad es 3/8.

Si X denota la *variable aleatoria* que indica la cantidad de niños en una familia con tres hijos, la distribución de probabilidad es la que se muestra en la tabla 6.2.

Tabla 6.2

Cantidad de niños (hombres) X	0	1	2	3
Probabilidad $p(X)$	1/8	3/8	3/8	1/8

6.11 Graficar la distribución del problema 6.10.

SOLUCIÓN

La gráfica puede representarse ya sea como en la figura 6.7 o como en la figura 6.8. Obsérvese que en la figura 6.8 la suma de las áreas de los rectángulos es 1; en esta figura llamada *histograma de probabilidad*, X es considerada como una variable continua, aunque en realidad sea discreta, procedimiento que suele resultar útil. Por otro lado, la figura 6.7 se emplea cuando no se desea considerar a la variable como variable continua.

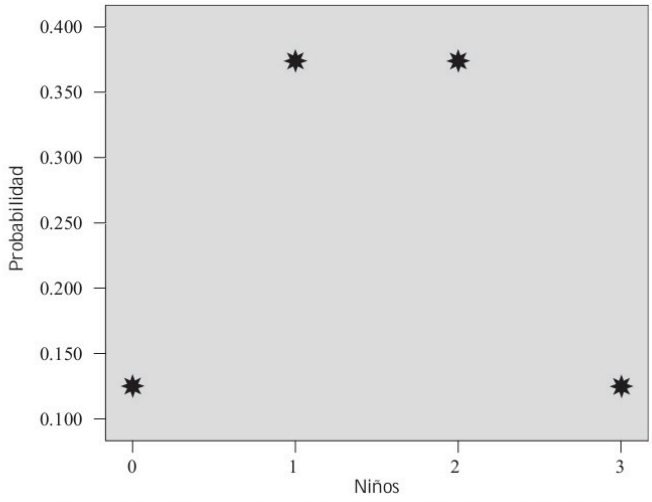


Figura 6-7 SPSS, gráfica de la distribución de probabilidad.

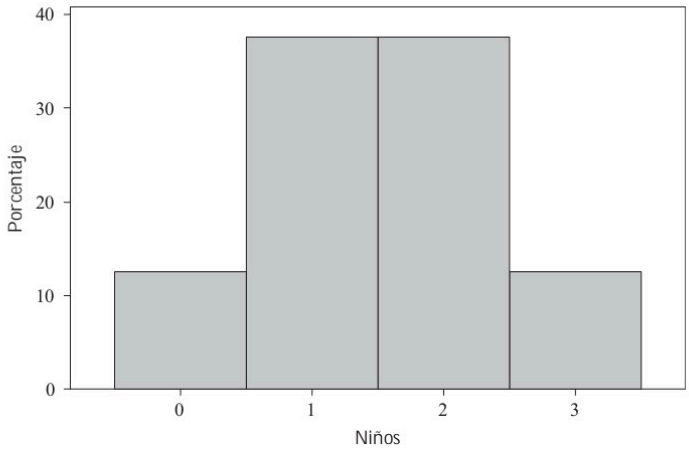


Figura 6-8 MINITAB, histograma de probabilidad.

6.12 Una variable aleatoria continua X, que toma valores sólo entre 0 y 5, tiene una función de probabilidad dada por $p(X) = \begin{cases} 0.2, & 0 < X < 5 \\ 0, & \text{si no es así} \end{cases}$. La gráfica se muestra en la figura 6.9.

a) Verificar que es una función de densidad.

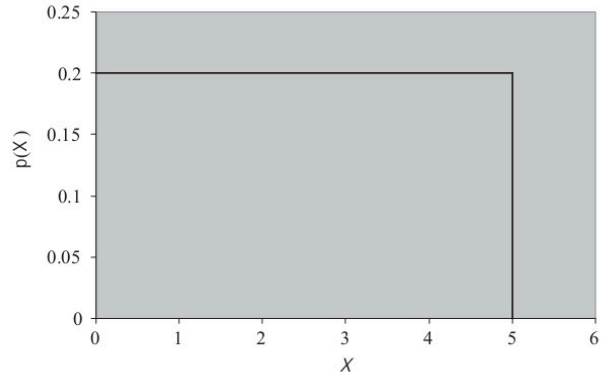


Figura 6-9 Función de densidad de probabilidad para la variable X.

b) Encontrar y graficar $\Pr\{2.5 < X < 4.0\}$.

SOLUCIÓN

a) La función $p(X)$ es siempre ≥ 0 y el área total bajo la gráfica de $p(X)$ es $5 \times 0.2 = 1$, ya que tiene forma rectangular con 0.2 de ancho y 5 de largo (ver figura 6-9).

b) La probabilidad $\Pr\{2.5 < X < 4.0\}$ se muestra en la figura 6.10.

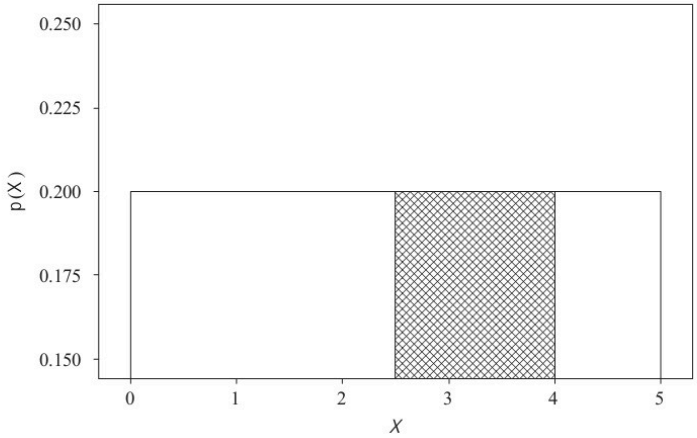


Figura 6-10 La probabilidad $\Pr\{2.5 < X < 4.0\}$ aparece como un área sombreada.

El área rectangular, $\Pr\{2.5 < X < 4.0\}$ es $(4 - 2.5) \times 0.2 = 0.3$.

ESPERANZA MATEMÁTICA

6.13 Se compra un boleto para una rifa con el que se puede ganar \$5 000 como primer premio o \$2 000 como segundo premio, siendo las probabilidades 0.001 y 0.003, respectivamente. ¿Cuál sería el precio justo a pagar por un boleto?

SOLUCIÓN

La esperanza es (\$5 000)(0.001) + (\$2 000)(0.003) = \$5 + \$6 = \$11, que es el precio justo a pagar.

6.14 En una inversión de negocios hay una probabilidad de 0.6 de obtener como ganancia \$300 y una probabilidad de 0.4 de perder \$100. Determinar la esperanza.

SOLUCIÓN

La esperanza es (\$300)(0.6) + (−\$100)(0.4) = \$180 − \$40 = \$140.

6.15 Encontrar: *a*) $E(X)$, *b*) $E(X^2)$ y *c*) $E[(X - \bar{X})^2)]$, para la distribución de probabilidad que se muestra en la tabla 6.3.
d) Emplear EXCEL para dar la solución de los incisos *a*), *b*) y *c*).

Tabla 6.3

X	8	12	16	20	24
$p(X)$	1/8	1/6	3/8	1/4	1/12

SOLUCIÓN

a) $E(X) = \sum Xp(X) = (8)(\frac{1}{8}) + (12)(\frac{1}{6}) + (16)(\frac{3}{8}) + (20)(\frac{1}{4}) + (24)(\frac{1}{12}) = 16$; lo que representa la *media* de la distribución.

b) $E(X^2) = \sum X^2 p(X) = (8)^2(\frac{1}{8}) + (12)^2(\frac{1}{6}) + (16)^2(\frac{3}{8}) + (20)^2(\frac{1}{4}) + (24)^2(\frac{1}{12}) = 276$; lo que representa el *segundo momento* respecto al origen cero.

c) $E[(X - \bar{X})^2] = \sum (X - \bar{X})^2 p(X) = (8 - 16)^2(\frac{1}{8}) + (12 - 16)^2(\frac{1}{6}) + (16 - 16)^2(\frac{3}{8}) + (20 - 16)^2(\frac{1}{4}) + (24 - 16)^2(\frac{1}{12}) = 20$; lo que representa la *varianza* de la distribución.

d) En A1:E1 se ingresan los títulos, como se muestra. Los valores de X y los valores de probabilidad se ingresan en A2:B6. Los valores esperados de X se calculan en C2:C7. El valor esperado se da en C7. El segundo momento respecto al origen se calcula en D2:D7. El segundo momento aparece en D7. La varianza se calcula en E2:E7. La varianza se da en E7.

A	B	C	D	E
X	$P(X)$	$Xp(X)$	$X^{2}p(X)$	$(X - E(X))^{2*}p(X)$
8	0.125	1	8	8
12	0.166667	2	24	2.666666667
16	0.375	6	96	0
20	0.25	5	100	4
24	0.083333	2	48	5.333333333
		16	276	20

6.16 Una bolsa contiene 2 pelotas blancas y 3 pelotas negras. Cada una de cuatro personas, *A*, *B*, *C* y *D*, en este orden, extrae una pelota y no la devuelve a la bolsa. La primera que extraiga una pelota blanca recibirá \$10. Determinar las esperanzas de *A*, *B*, *C* y *D*.

SOLUCIÓN

Como sólo hay 3 pelotas negras, alguna de las personas deberá ganar en el primer intento. Sean A, B, C y D los eventos "A gana", "B gana", "C gana" y "D gana", respectivamente.

$$\Pr\{A \text{ gane}\} = \Pr\{A\} = \frac{2}{3+2} = \frac{2}{5}$$

Por lo tanto, la esperanza de $A = \frac{2}{5}(\$10) = \$4$.

$$\Pr\{A \text{ pierda y } B \text{ gane}\} = \Pr\{\bar{A}B\} = \Pr\{\bar{A}\}\ \Pr\{B|\bar{A}\} = \left(\frac{3}{5}\right)\left(\frac{2}{4}\right) = \frac{3}{10}$$

Por lo tanto, la esperanza de $B = \$3$.

$$\Pr\{A \text{ y } B \text{ pierdan y } C \text{ gane}\} = \Pr\{\bar{A}\bar{B}C\} = \Pr\{\bar{A}\}\ \Pr\{\bar{B}|\bar{A}\}\ \Pr\{C\bar{A}\bar{B}\} = \left(\frac{3}{5}\right)\left(\frac{2}{4}\right)\left(\frac{2}{3}\right) = \frac{1}{5}$$

Por lo tanto, la esperanza de $C = \$2$.

$$\begin{aligned}\Pr\{A, B \text{ y } C \text{ pierdan y } D \text{ gane}\} &= \Pr\{\bar{A}\bar{B}\bar{C}\bar{D}\} \\ &= \Pr\{\bar{A}\}\ \Pr\{\bar{B}|\bar{A}\}\ \Pr\{\bar{C}|\bar{A}\bar{B}\}\ \Pr\{D|\bar{A}\bar{B}\bar{C}\} \\ &= \left(\frac{3}{5}\right)\left(\frac{2}{4}\right)\left(\frac{1}{3}\right)\left(\frac{1}{1}\right) = \frac{1}{10}\end{aligned}$$

Por lo tanto, la esperanza de $D = \$1$.

Comprobación: $\$4 + \$3 + \$2 + \$1 = \$10$ y $\frac{2}{5} + \frac{3}{10} + \frac{1}{5} + \frac{1}{10} = 1$.

PERMUTACIONES

6.17 ¿De cuántas maneras se pueden acomodar en línea 5 canicas de colores diferentes?

SOLUCIÓN

Hay que ordenar las cinco canicas en cinco posiciones: $- - - - -$. La primera posición puede ser ocupada por cualquiera de las 5 canicas (es decir, hay 5 maneras de ocupar la primera posición). Hecho esto, hay 4 maneras de ocupar la segunda posición; a continuación hay 3 maneras de ocupar la tercera posición; 2 maneras de ocupar la cuarta posición y, por último, sólo una manera de ocupar la última posición. Por lo tanto:

$$\text{Número de maneras en que se pueden colocar las cinco canicas en línea} = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5! = 120$$

En general,

$$\text{Número de maneras en las que se pueden colocar } n \text{ objetos diferentes en línea} = n(n-1)(n-2)\cdots 1 = n!$$

A esto se le conoce como el número de *permutaciones* de n objetos diferentes tomados de n en n y se denota ${}_nP_n$.

6.18 ¿De cuántas maneras se pueden sentar 10 personas en una banca en la que sólo hay 4 asientos disponibles?

SOLUCIÓN

Hay 10 maneras de ocupar el primer asiento; hecho esto, hay 9 maneras de ocupar el segundo asiento; 8 maneras de ocupar el tercer asiento, y 7 maneras de ocupar el cuarto asiento. Por lo tanto:

$$\text{Número de ordenaciones de 10 personas tomadas de 4 en 4} = 10 \cdot 9 \cdot 8 \cdot 7 = 5\,040$$

En general,

$$\text{Número de ordenaciones de } n \text{ objetos diferentes tomados de } r \text{ en } r = n(n-1)\cdots(n-r+1)$$

A esto también se le conoce como el número de *permutaciones* de n objetos diferentes tomados de r en r y se denota ${}_nP_r$, $P(n, r)$ o $P_{n,r}$. Obsérvese que cuando $r = n$, ${}_nP_n = n!$, como en el problema 6.17.

6.19 Evaluar: *a*) ${}_8P_3$, *b*) ${}_6P_4$, *c*) ${}_{15}P_1$ y ${}_3P_3$, y *e*) los incisos del *a*) al *d*) empleando EXCEL.

SOLUCIÓN

a) ${}_8P_3 = 8 \cdot 7 \cdot 6 = 336$, *b*) ${}_6P_4 = 6 \cdot 5 \cdot 4 \cdot 3 = 360$, *c*) ${}_{15}P_1 = 15$, y *d*) ${}_3P_3 = 3 \cdot 2 \cdot 1 = 6$

e) =PERMUTACIONES(8, 3) = 336 =PERMUTACIONES(6, 4) = 360
=PERMUTACIONES(15, 1) = 15 =PERMUTACIONES(3, 3) = 6

6.20 Se desea sentar en hilera a 5 hombres y 4 mujeres de manera que las mujeres ocupen los lugares pares. ¿De cuántas maneras es posible hacer esto?

SOLUCIÓN

Los hombres se pueden sentar de ${}_5P_5$ maneras y las mujeres de ${}_4P_4$ maneras. A cada acomodo de los hombres se le puede hacer corresponder un acomodo de las mujeres. Por lo tanto, la cantidad de acomodos es ${}_5P_5 \cdot {}_4P_4 = 5!4! = (120)(24) = 2\,880$.

6.21 ¿Cuántos números de cuatro dígitos se pueden formar con los 10 dígitos 0, 1, 2, 3,..., 9, si: *a*) puede haber repeticiones, *b*) no puede haber repeticiones y *c*) no puede haber repeticiones y el último dígito debe ser cero?

SOLUCIÓN

a) El primer dígito puede ser cualquiera de 9 dígitos (ya que no puede ser 0). El segundo, tercero y cuarto dígitos pueden ser uno cualquiera de 10. Entonces, se pueden formar $9 \cdot 10 \cdot 10 \cdot 10 = 9\,000$ números.

b) El primer dígito puede ser cualquiera de 9 dígitos (cualquiera menos el 0).
El segundo dígito puede ser cualquiera de 9 dígitos (cualquiera menos el usado como primer dígito).
El tercer dígito puede ser cualquiera de 8 dígitos (cualquiera menos los usados como los dos primeros dígitos).
El cuarto dígito puede ser cualquiera de 7 dígitos (cualquiera menos los usados como los primeros tres dígitos).
De manera que se pueden formar $9 \cdot 9 \cdot 8 \cdot 7 = 4\,536$ números.

Otro método

El primer dígito puede ser uno cualquiera de 9 dígitos y los tres restantes se pueden escoger de ${}_9P_3$ maneras. Por lo tanto, se pueden formar $9 \cdot {}_9P_3 = 9 \cdot 9 \cdot 8 \cdot 7 = 4\,536$ números.

c) El primer dígito se puede formar de 9 maneras, el segundo de 8 maneras y el tercero de 7 maneras. Por lo tanto, se pueden formar $9 \cdot 8 \cdot 7 = 504$ números.

Otro método

El primer dígito se puede formar de 9 maneras y los siguientes dos dígitos en ${}_9P_2$ maneras. Por lo tanto, se pueden encontrar $9 \cdot {}_8P_2 = 9 \cdot 8 \cdot 7 = 504$ números.

6.22 En un librero se van a acomodar cuatro libros diferentes de matemáticas, 6 libros diferentes de física y 2 libros diferentes de química. ¿Cuántos son los acomodos posibles si: *a*) los libros de cada materia tienen que estar juntos y *b*) sólo los libros de matemáticas tienen que estar juntos?

SOLUCIÓN

a) Los libros de matemáticas se pueden ordenar entre ellos de ${}_4P_4 = 4!$ maneras, los libros de física de ${}_6P_6 = 6!$ maneras, los libros de química de ${}_2P_2 = 2!$ maneras y los tres grupos de ${}_3P_3 = 3!$ maneras. Por lo tanto, el número de acomodos que se busca es $= 4!\ 6!\ 2!\ 3! = 207\,360$.

b) Considere a los 4 libros de matemáticas como un solo libro. Entonces se tienen 9 libros que se pueden acomodar de ${}_9P_9 = 9!$ maneras. En todas estas maneras, los libros de matemáticas están juntos. Pero los libros de matemáticas, entre ellos, se pueden acomodar de ${}_4P_4 = 4!$ maneras. Por lo tanto, el número de acomodos buscado es $= 9!\,4! = 8\,709\,120$.

6.23 Cinco canicas rojas, 2 canicas blancas y 3 azules están ordenadas en línea. Si las canicas de un mismo color no se distinguen unas de otras, ¿cuántas ordenaciones distintas se pueden tener? Para evaluar esta expresión usar la función de EXCEL definida como =MULTINOMIAL.

SOLUCIÓN

Supóngase que existen P ordenaciones diferentes. Multiplicando P por el número de maneras en las que se pueden ordenar: *a*) las 5 canicas rojas entre sí, *b*) las 2 canicas blancas entre sí y *c*) las 3 canicas azules entre sí (es decir, multiplicando P por 5! 2! 3!), se obtiene el número de maneras en que se pueden ordenar las 10 canicas si son distinguibles (es decir, 10!). Por lo tanto,

$$(5!2!3!)P = 10! \qquad \text{y} \qquad P = \frac{10!}{5!2!3!}$$

En general, el número de ordenaciones de n objetos de los cuales n_1 son iguales, n_2 son iguales, ..., n_k son iguales es

$$\frac{n!}{n_1!n_2!\cdots n_k!}$$

donde $n_1 + n_2 + \cdots + n_k = n$.

Con la función de EXCEL definida como =MULTINOMIAL(5,2,3) se obtiene 2 520.

6.24 ¿De cuántas maneras pueden sentarse 7 personas a una mesa redonda si: *a*) las 7 se pueden sentar en cualquier lugar y *b*) 2 determinadas personas no pueden sentarse juntas?

SOLUCIÓN

a) Se escoge una de las personas para sentarla en cualquier lugar. Entonces, las 6 personas restantes se pueden sentar de $6! = 720$ maneras, que es el total de maneras de acomodar a 7 personas en una mesa redonda.

b) Considérese como una sola persona a las dos personas que no se pueden sentar juntas. Entonces, quedan 6 personas en total que se pueden acomodar de 5! maneras. Pero las dos personas consideradas como una sola, entre ellas, se pueden acomodar de 2! maneras. Por lo tanto, la cantidad de maneras en que se pueden acomodar 6 personas en una mesa redonda sentando 2 personas juntas es $5!2! = 240$.

Entonces, empleando el inciso *a*), el total de maneras en las que 7 personas se pueden sentar a una mesa redonda, de manera que 2 determinadas personas no se sienten juntas $= 720 - 240 = 480$ maneras.

COMBINACIONES

6.25 ¿De cuántas maneras pueden colocarse 10 objetos en dos grupos, uno de 4 y otro de 6 objetos?

SOLUCIÓN

Esto es lo mismo que el número de ordenaciones de 10 objetos de los cuales 4 son iguales entre sí y 6 son iguales entre sí. De acuerdo con el problema 6.23, esto es

$$\frac{10!}{4!6!} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{4!} = 210$$

Este problema es equivalente a hallar de cuántas maneras se pueden tomar 4 de 10 objetos (o bien 6 de 10 objetos) sin importar el orden.

En general, el número de maneras en que se pueden seleccionar r de n objetos, a lo que se le llama el número de *combinaciones* de n cosas tomadas de r en r, se denota $\binom{n}{r}$ y está dado por

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{{}_nP_r}{r!}$$

6.26 Evaluar: *a*) $\binom{7}{4}$, *b*) $\binom{6}{5}$, *c*) $\binom{4}{4}$ y *d*) evaluar los incisos del *a*) al *c*) empleando EXCEL.

SOLUCIÓN

a)
$$\binom{7}{4} = \frac{7!}{4!\,3!} = \frac{7 \cdot 6 \cdot 5 \cdot 4}{4!} = \frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} = 35$$

b)
$$\binom{6}{5} = \frac{6!}{5!\,1!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{5!} = 6 \qquad \text{o bien} \qquad \binom{6}{5} = \binom{6}{1} = 6$$

c) $\binom{4}{4}$ es el número de maneras en que se pueden tomar todos los cuatro objetos, y sólo hay una manera, por lo que $\binom{4}{4} = 1$. Obsérvese que formalmente

$$\binom{4}{4} = \frac{4!}{4!\,0!} = 1$$

si definimos $0! = 1$.

d) =COMBIN(7, 4) da 35, =COMBIN(6, 5) da 6 y =COMBIN(4, 4) da 1.

6.27 ¿De cuántas maneras puede formarse de un grupo de 9 personas un comité de 5 personas?

SOLUCIÓN

$$\binom{9}{5} = \frac{9!}{5!\,4!} = \frac{9 \cdot 8 \cdot 7 \cdot 6 \cdot 5}{5!} = 126$$

6.28 Con 5 matemáticos y 7 físicos hay que formar un comité que conste de 2 matemáticos y 3 físicos. ¿De cuántas maneras se puede formar este comité si: *a*) puede incluirse a cualquiera de los matemáticos y a cualquiera de los físicos, *b*) hay uno de los físicos que tiene que formar parte del comité y *c*) hay dos de los matemáticos que no pueden formar parte del comité?

SOLUCIÓN

a) Dos matemáticos de 5 se pueden seleccionar de $\binom{5}{2}$ formas y 3 físicos de 7 se pueden seleccionar de $\binom{7}{3}$ formas. Así que las maneras en que se puede seleccionar el comité son

$$\binom{5}{2} \cdot \binom{7}{3} = 10 \cdot 35 = 350$$

b) Dos matemáticos de 5 se pueden seleccionar de $\binom{5}{2}$ maneras y 2 físicos de 6 se pueden seleccionar de $\binom{6}{2}$ maneras. Así que las maneras en que se puede seleccionar el comité son

$$\binom{5}{2} \cdot \binom{6}{2} = 10 \cdot 15 = 150$$

c) Dos matemáticos de 3 se pueden seleccionar de $\binom{3}{2}$ maneras y 3 físicos de 7 se pueden seleccionar de $\binom{7}{3}$ maneras. Así que las maneras en que se puede seleccionar el comité son

$$\binom{3}{2} \cdot \binom{7}{3} = 3 \cdot 35 = 105$$

6.29 Una niña tiene 5 flores que son todas distintas. ¿Cuántos ramos puede formar?

SOLUCIÓN

Cada flor puede tratarse de dos maneras; puede ser elegida para el ramo o puede no ser elegida para el ramo. Como a cada una de estas dos maneras de tratar a la flor le corresponden 2 maneras de tratar a cada una de las otras flores, el número de maneras en que se puede tratar a las 5 flores es $= 2^5$. Pero estas 2^5 maneras comprenden el caso en que no se elija ninguna de las flores. Por lo tanto, la cantidad de ramos que pueden formarse es $= 2^5 - 1 = 31$.

Otro método

La niña puede elegir 1 de 5 flores, 2 de 5 flores, ..., 5 de 5 flores. Por lo tanto, el número de ramos que puede formar es

$$\binom{5}{1} + \binom{5}{2} + \binom{5}{3} + \binom{5}{4} + \binom{5}{5} = 5 + 10 + 10 + 5 + 1 = 31$$

En general, para todo número entero positivo n,

$$\binom{n}{1} + \binom{n}{2} + \binom{n}{3} + \cdots + \binom{n}{n} = 2^n - 1$$

6.30 Con 7 consonantes y 5 vocales ¿cuántas palabras con 4 consonantes distintas y 3 vocales distintas pueden formarse? No importa que las palabras no tengan significado.

SOLUCIÓN

Las cuatro consonantes distintas pueden elegirse de $\binom{7}{4}$ maneras, las tres vocales distintas pueden elegirse de $\binom{5}{3}$ maneras, y estas 7 letras (4 consonantes y 3 vocales) pueden ordenarse de ${}_7P_7 = 7!$ maneras. Por lo tanto, el número de palabras es

$$\binom{7}{4} \cdot \binom{5}{3} \cdot 7! = 35 \cdot 10 \cdot 5\,040 = 1\,764\,000$$

APROXIMACIÓN DE STIRLING PARA *n!*

6.31 Evaluar 50!

SOLUCIÓN

Cuando n es grande se tiene $n! \approx \sqrt{2\pi n}\, n^n e^{-n}$; por lo tanto,

$$50! \approx \sqrt{2\pi(50)}\, 50^{50} e^{-50} = S$$

Para evaluar S se usan logaritmos base 10. Por lo tanto,

$$\begin{aligned} \log S = \log(\sqrt{100\pi}\, 50^{50} e^{-50}) &= \tfrac{1}{2}\log 100 + \tfrac{1}{2}\log \pi + 50 \log 50 - 50 \log e \\ &= \tfrac{1}{2}\log 100 + \tfrac{1}{2}\log 3.142 + 50 \log 50 - 50 \log 2.718 \\ &= \tfrac{1}{2}(2) + \tfrac{1}{2}(0.4972) + 50(1.6990) - 50(0.4343) = 64.4846 \end{aligned}$$

de lo que se encuentra que $S = 3.05 \times 10^{64}$, número que tiene 65 dígitos.

PROBABILIDAD Y ANÁLISIS COMBINATORIO

6.32 Una caja contiene 8 pelotas rojas, 3 blancas y 9 azules. Si se extraen 3 pelotas en forma aleatoria, determinar la probabilidad de que: *a*) las 3 sean rojas, *b*) las 3 sean blancas, *c*) 2 sean rojas y 1 sea blanca, *d*) por lo menos 1 sea blanca, *e*) se extraiga una de cada color y *f*) se extraigan en el orden roja, blanca, azul.

SOLUCIÓN

a) **Primer método**

Sean R_1, R_2 y R_3 los eventos "pelota roja en la primera extracción", "pelota roja en la segunda extracción", "pelota roja en la tercera extracción", respectivamente. Entonces $R_1R_2R_3$ denota el evento de que las tres pelotas extraídas sean rojas.

$$\Pr\{R_1R_2R_3\} = \Pr\{R_1\}\ \Pr\{R_2|R_1\}\ \Pr\{R_3|R_1R_2\} = \left(\frac{8}{20}\right)\left(\frac{7}{19}\right)\left(\frac{6}{18}\right) = \frac{14}{285}$$

Segundo método

$$\text{Probabilidad buscada} = \frac{\text{maneras de seleccionar 3 de 8 pelotas rojas}}{\text{maneras de seleccionar 3 de 20 pelotas}} = \frac{\binom{8}{3}}{\binom{20}{3}} = \frac{14}{285}$$

b) Empleando el segundo método del inciso *a*),

$$\Pr\{\text{las 3 sean blancas}\} = \frac{\binom{3}{3}}{\binom{20}{3}} = \frac{1}{1\,140}$$

También se puede usar el primer método del inciso *a*)

c) $$\Pr\{\text{2 sean rojas y 1 sea blanca}\} = \frac{\binom{\text{maneras de seleccionar}}{\text{2 de 8 pelotas rojas}}\binom{\text{maneras de seleccionar}}{\text{1 de 3 pelotas blancas}}}{\text{maneras de seleccionar 3 de 20 pelotas}} = \frac{\binom{8}{2}\binom{3}{1}}{\binom{20}{3}} = \frac{7}{95}$$

d) $$\Pr\{\text{ninguna sea blanca}\} = \frac{\binom{17}{3}}{\binom{20}{3}} = \frac{34}{57} \qquad \text{de manera que} \qquad \Pr\{\text{por lo menos 1 sea blanca}\} = 1 - \frac{34}{57} = \frac{23}{57}$$

e) $$\Pr\{\text{1 de cada color}\} = \frac{\binom{8}{1}\binom{3}{1}\binom{9}{1}}{\binom{20}{3}} = \frac{18}{95}$$

f) Empleando el inciso *e*),

$$\Pr\{\text{extraer las pelotas en el orden rojo, blanco, azul}\} = \frac{1}{3!}\Pr\{\text{1 de cada color}\} = \frac{1}{6}\left(\frac{18}{95}\right) = \frac{3}{95}$$

Otro método

$$\Pr\{R_1W_2B_2\} = \Pr\{R_1\}\Pr\{W_2|R_1\}\Pr\{B_3|R_1W_2\} = \left(\frac{8}{20}\right)\left(\frac{3}{19}\right)\left(\frac{9}{18}\right) = \frac{3}{95}$$

6.33 De una baraja de 52 cartas bien barajadas se extraen 5 cartas. Encontrar la probabilidad de que: *a*) 4 sean ases; *b*) 4 sean ases y 1 sea rey; *c*) 3 sean dieces y 2 sean sotas; *d*) sean 9, 10, sota, reina y rey en cualquier orden; *e*) 3 sean de un palo y 2 de otro palo, y *f*) se obtenga por lo menos 1 as.

SOLUCIÓN

a) $$\Pr\{\text{4 ases}\} = \binom{4}{4} \cdot \frac{\binom{48}{1}}{\binom{52}{5}} = \frac{1}{54\,145}$$

b) $$\Pr\{\text{4 ases y 1 rey}\} = \binom{4}{4} \cdot \frac{\binom{4}{1}}{\binom{52}{5}} = \frac{1}{649\,740}$$

c) $$\Pr\{\text{3 sean dieces y 2 sean sotas}\} = \binom{4}{3} \cdot \frac{\binom{4}{2}}{\binom{52}{5}} = \frac{1}{108\,290}$$

d) $$\Pr\{\text{sean 9, 10, sota, reina y rey en cualquier orden}\} = \frac{\binom{4}{1} \cdot \binom{4}{1} \cdot \binom{4}{1} \cdot \binom{4}{1} \cdot \binom{4}{1}}{\binom{52}{5}} = \frac{64}{162\,435}$$

e) Como hay 4 maneras de elegir el primer palo y 3 maneras de elegir el segundo palo,

$$\Pr\{\text{3 sean de un palo y 2 de otro palo}\} = \frac{4\binom{13}{3} \cdot 3\binom{13}{2}}{\binom{52}{5}} = \frac{429}{4\,165}$$

f) $$\Pr\{\text{ningún as}\} = \frac{\binom{48}{5}}{\binom{52}{5}} = \frac{35\,673}{54\,145} \qquad \text{y} \qquad \Pr\{\text{por lo menos 1 as}\} = 1 - \frac{35\,673}{54\,145} = \frac{18\,482}{54\,145}$$

6.34 Determinar la probabilidad de tener 3 seises en cinco lanzamientos de un dado.

SOLUCIÓN

Los lanzamientos del dado se representarán por 5 espacios: $_\ _\ _\ _\ _$. En cada espacio se tendrá el evento 6 o el evento no-6 ($\bar{6}$); por ejemplo se pueden tener tres 6 y dos no-6 en esta forma $6\ 6\ \bar{6}\ 6\ \bar{6}$ o en esta forma $6\ \bar{6}\ 6\ \bar{6}\ 6$, etcétera.

Ahora la probabilidad de un evento, como, por ejemplo, $6\ 6\ \bar{6}\ 6\ \bar{6}$ es

$$\Pr\{6\,6\,\bar{6}\,6\,\bar{6}\} = \Pr\{6\}\Pr\{6\}\ \Pr\{\bar{6}\}\ \Pr\{6\}\ \Pr\{\bar{6}\} = \frac{1}{6}\cdot\frac{1}{6}\cdot\frac{5}{6}\cdot\frac{1}{6}\cdot\frac{5}{6} = \left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2$$

De igual manera, $\Pr\{6\bar{6}\,6\bar{6}6\} = (\frac{1}{6})^3(\frac{5}{6})^2$, etc., para todos los eventos en los que hay tres 6 y dos no-6. Pero de estos eventos hay $\binom{5}{3} = 10$ y estos eventos son mutuamente excluyentes; por lo tanto, la probabilidad buscada es

$$\Pr\{6\,6\,\bar{6}\,6\,\bar{6} \text{ o } 6\,\bar{6}\,6\,\bar{6}\,6 \text{ o etc.}\} = \binom{5}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 = \frac{125}{3\,888}$$

En general si $q = \Pr\{E\}$ y $q = \Pr\{\bar{E}\}$, entonces empleando el razonamiento anterior, la probabilidad de en N ensayos obtener exactamente X veces E es $\binom{N}{X}p^X q^{N-X}$.

6.35 En una fábrica se encuentra que en promedio 20% de los tornillos producidos con una máquina están defectuosos. Si se toman aleatoriamente 10 tornillos producidos con esta máquina en un día, encontrar la probabilidad de que: *a*) exactamente 2 estén defectuosos, *b*) 2 o más estén defectuosos y *c*) más de 5 estén defectuosos.

SOLUCIÓN

a) Empleando un razonamiento similar al empleado en el problema 6.34,

$$\Pr\{\text{2 tornillos defectuosos}\} = \binom{10}{2}(0.2)^2(0.8)^8 = 45(0.04)(0.1678) = 0.3020$$

b) $\Pr\{\text{2 o más tornillos defectuosos}\} = 1 - \Pr\{\text{0 tornillos defectuosos}\} - \Pr\{\text{1 tornillo defectuoso}\}$

$$= 1 - \binom{10}{0}(0.2)^0(0.8)^{10} - \binom{10}{1}(0.2)^1(0.8)^9$$
$$= 1 - (0.8)^{10} - 10(0.2)(0.8)^9$$
$$= 1 - 0.1074 - 0.2684 = 0.6242$$

c) $\Pr\{\text{más de 5 tornillos defectuosos}\} = \Pr\{6 \text{ tornillos defectuosos}\} + \Pr\{7 \text{ tornillos defectuosos}\}$
$+ \Pr\{8 \text{ tornillos defectuosos}\} + \Pr\{9 \text{ tornillos defectuosos}\}$
$+ \Pr\{10 \text{ tornillos defectuosos}\}$
$= \binom{10}{6}(0.2)^6(0.8)^4 + \binom{10}{7}(0.2)^7(0.8)^3 + \binom{10}{8}(0.2)^8(0.8)^2$
$+ \binom{10}{9}(0.2)^9(0.8) + \binom{10}{10}(0.2)^{10}$
$= 0.00637$

6.36 Si en el problema 6.35 se tomaron 1 000 muestras de 10 tornillos cada una, ¿en cuántas de estas muestras se espera encontrar: *a*) exactamente 2 tornillos defectuosos, *b*) 2 o más tornillos defectuosos y *c*) más de 5 tornillos defectuosos?

SOLUCIÓN

a) La cantidad esperada es = (1 000)(0.3020) = 302, de acuerdo con el problema 6.35*a*).
b) La cantidad esperada es = (1 000)(0.6242) = 624, de acuerdo con el problema 6.35*b*).
c) La cantidad esperada es = (1 000)(0.00637) = 6, de acuerdo con el problema 6.35*c*).

DIAGRAMAS DE EULER O DE VENN Y PROBABILIDAD

6.37 En la figura 6.11 se muestra cómo representar el espacio muestral de cuatro lanzamientos de una moneda y los eventos E_1, obtener exactamente dos caras y dos cruces, y E_2, obtener lo mismo en el primero y en el último lanzamiento. Ésta es una manera de representar diagramas de Venn y eventos en una hoja de cálculo.

	espacio	muestras		evento E1	evento E2
h	h	h	h		Y
h	h	h	t		
h	h	t	h		Y
h	h	t	t	X	
h	t	h	h		Y
h	t	h	t	X	
h	t	t	h	X	Y
h	t	t	t		
t	h	h	h		
t	h	h	t	X	Y
t	h	t	h	X	
t	h	t	t		Y
t	t	h	h	X	
t	t	h	t		Y
t	t	t	h		
t	t	t	t		Y

Figura 6-11 EXCEL, representación del espacio muestral y de los eventos E_1 y E_2.

Debajo de E_1 se han marcado con una X los casos en los que se da el evento E_1 y debajo de E_2 se han marcado con una Y los casos en los que se da el evento E_2.

a) Dar los casos que pertenecen a $E_1 \cap E_2$ y $E_1 \cup E_2$.
b) Dar las probabilidades $\Pr\{E_1 \cap E_2\}$ y $\Pr\{E_1 \cup E_2\}$.

SOLUCIÓN

a) Los casos que pertenecen a $E_1 \cap E_2$ son los que tienen X y Y. Por lo tanto, $E_1 \cap E_2$ consta de los casos htth y thht. Los casos que pertenecen a $E_1 \cup E_2$ son los que tienen X, Y o X y Y. Los casos que pertenecen a $E_1 \cup E_2$ son: hhhh, hhth, hhtt, hthh, htht, htth, thht, thth, thtt, tthh, ttht y tttt.
b) $\Pr\{E_1 \cap E_2\} = 2/16 = 1/8$ o 0.125. $\Pr\{E_1 \cup E_2\} = 12/16 = 3/4$ o 0.75.

6.38 Usando un espacio muestral y diagramas de Venn, mostrar que

a) $\Pr\{A \cup B\} = \Pr\{A\} + \Pr\{B\} - \Pr\{A \cap B\}$

b) $\Pr\{A \cup B \cup C\} = \Pr\{A\} + \Pr\{B\} + \Pr\{C\} - \Pr\{A \cap B\} - \Pr\{B \cap C\} - \Pr\{A \cap C\} + \Pr\{A \cap B \cap C\}$

SOLUCIÓN

a) La unión no mutuamente excluyente $A \cup B$ se puede expresar como la unión mutuamente excluyente de $A \cap \bar{B}$, $B \cap \bar{A}$, y $A \cap B$.

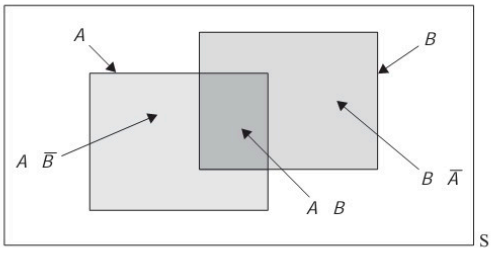


Figura 6-12 Una unión expresada como unión disyunta.

$$\Pr\{A \cup B\} = \Pr\{A \cap \bar{B}\} + \Pr\{B \cap \bar{A}\} + \Pr\{A \cap B\}$$

Ahora en el lado derecho de la ecuación se suma y se resta $\Pr\{A \cap B\}$.

$$\Pr\{A \cup B\} = \Pr\{A \cap \bar{B}\} + \Pr\{B \cap \bar{A}\} + \Pr\{A \cap B\} + [\Pr\{A \cap B\} - \Pr\{A \cap B\}]$$

Reordenando esta ecuación de la manera siguiente:

$$\Pr\{A \cup B\} = [\Pr\{A \cap \bar{B}\} + \Pr\{A \cap B\}] + [\Pr\{B \cap \bar{A}\} + Pr\{A \cap B\}] - \Pr\{A \cap B\}$$

$$\Pr\{A \cup B\} = \Pr\{A\} + \Pr\{B\} - \Pr\{A \cap B\}$$

b) En la figura 6-13, el evento A está compuesto por las regiones 1, 2, 3 y 6, el evento B está compuesto por las regiones 1, 3, 4 y 7, y el evento C está compuesto por las regiones 1, 2, 4 y 5.

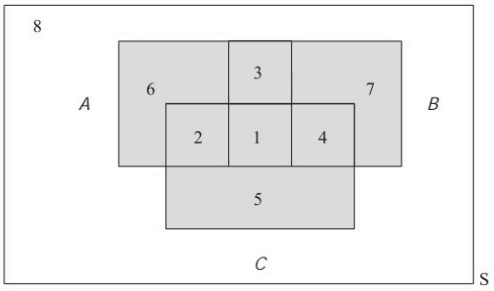


Figura 6-13 La unión no mutuamente excluyente de tres eventos, $A \cup B \cup C$.

El espacio muestral de la figura 6-13 está formado por 8 regiones mutuamente excluyentes. Estas 8 regiones se describen como sigue: la región 1 es $A \cap B \cap C$, la región 2 es $A \cap C \cap \bar{B}$, la región 3 es $A \cap B \cap \bar{C}$, la región 4 es $\bar{A} \cap C \cap B$, la región 5 es $\bar{A} \cap C \cap \bar{B}$, la región 6 es $A \cap \bar{C} \cap \bar{B}$, la región 7 es $\bar{A} \cap \bar{C} \cap B$ y la región 8 es $\bar{A} \cap \bar{C} \cap \bar{B}$.

La probabilidad $\Pr\{A \cup B \cup C\}$ se expresa como la probabilidad de las 7 regiones mutuamente excluyentes que forman $A \cup B \cup C$, como sigue:

$$\Pr\{A \cap B \cap C\} + \Pr\{A \cap C \cap \bar{B}\} + \Pr\{A \cap B \cap \bar{C}\} + \Pr\{\bar{A} \cap C \cap B\} + \Pr\{\bar{A} \cap C \cap \bar{B}\} + \Pr\{A \cap \bar{C} \cap \bar{B}\} + \Pr\{\bar{A} \cap \bar{C} \cap B\}$$

Cada parte de esta ecuación puede reescribirse y toda completa simplificarse para obtener:

$$\Pr\{A \cup B \cup C\} = \Pr\{A\} + \Pr\{B\} + \Pr\{C\} - \Pr\{A \cap B\} - \Pr\{B \cap C\} - \Pr\{A \cap C\} + \Pr\{A \cap B \cap C\}$$

Por ejemplo, $\Pr\{\bar{A} \cap C \cap \bar{B}\}$ puede expresarse como

$$\Pr\{C\} - \Pr\{A \cap C\} - Pr\{B \cap C\} + \Pr\{A \cap B \cap C\}$$

6.39 En una entrevista a 500 adultos se les hizo una pregunta que constaba de tres partes: 1) ¿Tiene usted teléfono celular? 2) ¿Tiene un ipod? 3) ¿Tiene conexión a Internet? Los resultados se presentan a continuación (ninguno contestó que no a todas las preguntas).

Teléfono celular	329	teléfono celular e ipod	83
ipod	186	teléfono celular y conexión a Internet	217
conexión a Internet	295	ipod y conexión a Internet	63

Dar la probabilidad de los eventos siguientes:

a) que conteste sí a todas las preguntas, *b*) que tenga teléfono celular, pero no conexión a Internet, *c*) que tenga ipod, pero no teléfono celular, *d*) que tenga conexión a Internet, pero no ipod *e*) que tenga teléfono celular o conexión a Internet pero no ipod y *g*) que tenga teléfono celular, pero no ipod o conexión a Internet.

SOLUCIÓN

El evento A es que el entrevistado tenga teléfono celular, el evento B es que el entrevistado tenga ipod y el evento C es que el entrevistado tenga conexión a Internet.

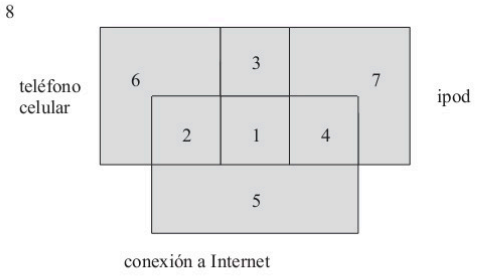


Figura. 6-14 Diagrama de Ven para el problema 6.39.

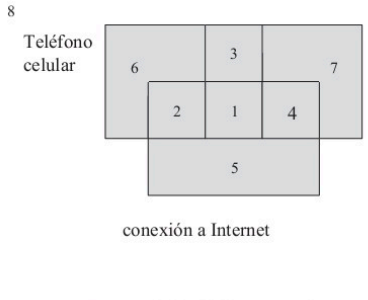
a) La probabilidad de que todos estén en la unión es 1, ya que ninguno respondió no a las tres partes. $\Pr\{A \cup B \cup C\}$ está dada por la expresión siguiente:

$$\Pr\{A\} + \Pr\{B\} + \Pr\{C\} - \Pr\{A \cap B\} - \Pr\{B \cap C\} - \Pr\{A \cap C\} + \Pr\{A \cap B \cap C\}$$

$$1 = 329/500 + 186/500 + 295/500 - 83/500 - 63/500 - 217/500 + \Pr\{A \cap B \cap C\}$$

Despejando $\Pr\{A \cap B \cap C\}$, se obtiene $1 - 447/500$ o bien $53/500 = 0.106$.

Antes de responder los demás incisos conviene llenar las regiones de la figura 6-14, como se muestra en la figura 6-15. La cantidad correspondiente a la región 2 es la cantidad en la región $A \cap C$ menos la cantidad en la región 1 o bien $217 - 53 = 164$. La cantidad correspondiente a la región 3 es la cantidad en la región $A \cap B$ menos la cantidad en la región 1 o bien $83 - 53 = 30$. La cantidad en la región 4 es la cantidad correspondiente a la región $B \cap C$ menos el número en la región 1 o bien $63 - 53 = 10$. La cantidad correspondiente a la región 5 es la cantidad en la región C menos la cantidad en las regiones 1, 2 y 4 o bien $295 - 53 - 164 - 10 = 68$. La cantidad correspondiente a la región 6 es la cantidad en la región A menos la cantidad en las regiones 1, 2 y 3 o bien $329 - 53 - 164 - 30 = 82$. La cantidad correspondiente a la región 7 es $186 - 53 - 30 - 10 = 93$.



Región	número
1	53
2	164
3	30
4	10
5	68
6	82
7	93
Total	500

Figura 6-15 *A*, *B* y *C* divididas en regiones mutuamente excluyentes.

b) regiones 3 y 6 o bien $30 + 82 = 112$ y la probabilidad es $112/500 = 0.224$.
c) regiones 4 y 7 o bien $10 + 93 = 103$ y la probabilidad es $103/500 = 0.206$.
d) regiones 2 y 5 o bien $164 + 68 = 232$ y la probabilidad es $232/500 = 0.464$.
e) regiones 2, 5 o bien 6 o bien $164 + 82 = 314$ y la probabilidad es $314/500 = 0.628$.
f) región 6 u 82 y la probabilidad es $82/500 = 0.164$.

PROBLEMAS SUPLEMENTARIOS

REGLAS FUNDAMENTALES DE LA PROBABILIDAD

6.40 Determinar o estimar la probabilidad p de cada uno de los eventos siguientes:

a) Al extraer de una baraja bien barajada una sola carta, obtener rey, as, sota de tréboles o rey de diamantes.
b) Se lanzan dos dados, una sola vez, y la suma de los puntos que aparecen en ellos resulte 8.
c) Encontrar un tornillo que no esté defectuoso si de 600 tornillos examinados, 12 estuvieron defectuosos.
d) Se lanzan dos dados una vez y la suma de los puntos resulte 7 u 11.
e) Al lanzar tres veces una moneda obtener cara por lo menos una vez.

6.41 Un experimento consiste en extraer sucesivamente tres cartas de una baraja bien barajada. Sea E_1 el evento "rey" en la primera extracción, E_2 el evento "rey" en la segunda extracción y E_3 el evento "rey" en la tercera extracción. Exprese en palabras el significado de:

a) $\Pr\{E_1\bar{E}_2\}$ *c*) $\bar{E}_1 + \bar{E}_2$ *e*) $\bar{E}_1\bar{E}_2\bar{E}_3$
b) $\Pr\{E_1 + E_2\}$ *d*) $\Pr\{E_3|E_1\bar{E}_2\}$ *f*) $\Pr\{E_1E_2 + \bar{E}_2E_3\}$

6.42 De una caja que contiene 10 canicas rojas, 30 blancas, 20 azules y 15 anaranjadas, se extrae una canica. Hallar la probabilidad de que la canica extraída sea: *a*) anaranjada o roja, *b*) ni azul ni roja, *c*) no azul, *d*) blanca y *e*) roja, blanca o azul.

6.43 De la caja del problema 6.42 se extraen sucesivamente dos canicas, devolviendo a la caja cada canica después de extraída. Encontrar la probabilidad de que: *a*) las dos sean blancas, *b*) la primera sea roja y la segunda sea blanca, *c*) ninguna sea anaranjada, *d*) sean rojas o blancas o las dos cosas (roja y blanca), *e*) la segunda no sea azul, *f*) la primera sea anaranjada, *g*) por lo menos una sea azul, *h*) cuando mucho una sea roja, *i*) la primera sea blanca, pero la segunda no, y *f*) sólo una sea roja.

6.44 Repetir el problema 6.43, pero suponiendo que una vez extraídas las canicas no se devuelven a la caja.

6.45 Encontrar la probabilidad de que al lanzar dos veces dos dados los puntos que se obtengan sumen 7: *a*) en el primer lanzamiento, *b*) en uno de los dos lanzamientos y *c*) en los dos lanzamientos.

6.46 De una baraja de 52 cartas, bien barajada, se extraen sucesivamente dos cartas. Encontrar la probabilidad de que: *a*) la primera carta extraída no sea un 10 de tréboles o un as, *b*) la primera carta sea un as, pero la segunda no, *c*) por lo menos una de las cartas sea un diamante, *d*) las cartas no sean de un mismo palo, *e*) no más de una de las cartas sea una figura (sota, reina o rey), *f*) la segunda carta no sea una figura, y *g*) la segunda carta no sea una figura dado que la primera sí fue una figura, *h*) las cartas sean figuras o espadas o ambas.

6.47 Una caja contiene papelillos numerados del 1 al 9. Si se extraen 3 papelillos de uno en uno, encontrar la probabilidad de que tengan números: 1) non, par, non o 2) par, non, par.

6.48 Las oportunidades de que A gane un partido de ajedrez contra B son 3:2. Si se van a jugar 3 partidos, ¿cuáles son las posibilidades: *a*) a favor de que A gane por lo menos dos de los tres partidos y *b*) en contra de que A pierda los dos primeros partidos contra B?

6.49 En un monedero hay dos monedas de plata y dos monedas de cobre, en otro monedero hay cuatro monedas de plata y 3 monedas de cobre. Si se toma al azar una moneda de uno de los dos monederos, ¿cuál es la probabilidad de que sea una moneda de plata?

6.50 La probabilidad de que en 25 años un hombre esté vivo es $\frac{3}{5}$ y la probabilidad de que en 25 años su esposa esté viva es $\frac{2}{3}$. Encontrar la probabilidad de que en 25 años: *a*) ambos estén vivos, *b*) sólo el hombre esté vivo, *c*) sólo la esposa esté viva y *d*) por lo menos uno esté vivo.

6.51 De 800 familias con cuatro hijos cada una, ¿qué porcentaje se espera que tenga: *a*) 2 niños y 2 niñas, *b*) por lo menos 1 niño, *c*) ninguna niña y *d*) cuando mucho 2 niñas? Supóngase que la probabilidad de niño y de niña es la misma.

DISTRIBUCIONES DE PROBABILIDAD

6.52 Si X es la variable aleatoria que indica la cantidad de niños en una familia con 4 hijos (ver problema 6.51): *a*) construir una tabla que dé la distribución de probabilidad de X y *b*) representar gráficamente la distribución de probabilidad del inciso *a*).

6.53 Una variable aleatoria continua que toma valores sólo entre $X = 2$ y $X = 8$, inclusive, tiene una función de densidad dada por $a(X + 3)$, donde a es una constante. *a*) Calcular a. Hallar *b*) $\Pr\{3 < X < 5\}$, *c*) $\Pr\{X \geq 4\}$ y *d*) $\Pr\{|X - 5| < 0.5\}$.

6.54 De una urna que contiene 4 canicas rojas y 6 blancas se extraen 3 canicas sin reemplazo. Si X es la variable aleatoria que indica la cantidad de canicas rojas extraídas: *a*) construir una tabla que muestre la distribución de probabilidad de X, y *b*) graficar la distribución.

6.55 *a*) Se lanzan 3 dados y $X =$ la suma de las tres caras que caen hacia arriba. Dar la distribución de probabilidad de X. *b*) Encontrar $\Pr\{7 \leq X \leq 11\}$.

ESPERANZA MATEMÁTICA

6.56 ¿Cuál es el precio justo a pagar en un juego en el que se pueden ganar \$25 con probabilidad 0.2 y \$10 con probabilidad 0.4?

6.57 Si llueve, un vendedor de paraguas gana \$30 diarios. Si no, pierde \$6 diarios. ¿Cuál es la esperanza si la probabilidad de que llueva es 0.3?

6.58 *A* y *B* juegan un partido en el que lanzan una moneda 3 veces. El primero que obtiene cara, gana el partido. Si *A* lanza primero la moneda y si el valor total de las apuestas es \$20, ¿con cuánto deberá contribuir cada uno para que el juego sea justo?

6.59 Dada la distribución de probabilidad de la tabla 6.4, hallar: *a*) $E(X)$, *b*) $E(X^2)$, *c*) $E[(X-\bar{X})^2]$, *d*) $E(X^3)$.

Tabla 6.4

X	-10	-20	30
$p(X)$	1/5	3/10	1/2

6.60 Dados los datos del problema 6.54, encontrar: *a*) la media, *b*) la varianza y *c*) la desviación estándar de la distribución de *X* e interpretar los resultados.

6.61 Una variable aleatoria toma el valor 1 con probabilidad *p* y el valor 0 con probabilidad $q = 1 - p$. Probar que: *a*) $E(X) = p$ y *b*) $E[(X-\bar{X})^2] = pq$.

6.62 Probar que: *a*) $E(2X+3) = 2E(X) + 3$ y *b*) $E[(X-\bar{X})^2] = E(X^2) - [E(X)]^2$.

6.63 En el problema 6.55, encontrar el valor esperado de *X*.

PERMUTACIONES

6.64 Evalar: *a*) ${}_4P_2$, *b*) ${}_7P_5$, y *c*) ${}_{10}P_3$. Dar la función de EXCEL para evaluar los incisos *a*), *b*) y *c*).

6.65 ¿Para qué valores de *n* es ${}_{n+1}P_3 = {}_nP_4$?

6.66 ¿De cuántas maneras se pueden sentar 5 personas en un sofá si el sofá sólo tiene 3 asientos?

6.67 ¿De cuántas maneras pueden ordenarse 7 libros en un librero si: *a*) pueden ordenarse como se desee, *b*) hay 3 libros que deben estar juntos y *c*) hay 2 libros que deben estar al final?

6.68 ¿Cuántos números de cinco dígitos diferentes pueden formarse con los dígitos 1, 2, 3, . . . , 9 si: *a*) el número debe ser non y *b*) si los dos primeros dígitos de cada número tienen que ser pares?

6.69 Resolver el problema 6.68 si se permiten dígitos repetidos.

6.70 ¿Cuántos números de tres dígitos pueden formarse con tres 4, cuatro 2 y dos 3?

6.71 ¿De cuántas maneras pueden sentarse a una mesa redonda 3 hombres y 3 mujeres si: *a*) sin ninguna restricción, *b*) hay dos mujeres que no pueden sentarse juntas y *c*) cada mujer debe estar entre dos hombres?

COMBINACIONES

6.72 Evaluar: *a*) $\binom{7}{3}$, *b*) $\binom{8}{4}$ y *c*) $\binom{10}{8}$. Dar la función de EXCEL para evaluar los incisos *a*), *b*) y *c*).

6.73 ¿Para qué valores de n se cumple que: $3\binom{n+1}{3} = 7\binom{n}{2}$?

6.74 ¿De cuántas maneras se pueden seleccionar 6 de 10 preguntas?

6.75 ¿Cuántos comités de 3 hombres y 4 mujeres pueden formarse a partir de un grupo de 8 hombres y 6 mujeres?

6.76 ¿De cuántas maneras pueden seleccionarse 2 hombres, 4 mujeres, 3 niños y 3 niñas de un grupo de 6 hombres, 8 mujeres, 4 niños y 5 niñas si: *a*) no hay ninguna restricción y *b*) hay un hombre y una mujer que tienen que seleccionarse?

6.77 ¿De cuántas maneras puede dividirse un grupo de 10 personas en: *a*) dos grupos de 7 y 3 personas y *b*) tres grupos de 4, 3, y 2 personas?

6.78 A partir de 5 profesionales de la estadística y 6 economistas, se va a formar un grupo que conste de 3 profesionales de la estadística y 2 economistas. ¿Cuántos comités diferentes pueden formarse si: *a*) no hay restricción alguna, *b*) hay 2 profesionales de la estadística que deben estar en el comité y *c*) hay un economista que no puede formar parte del comité?

6.79 Encontrar la cantidad de: *a*) combinaciones y *b*) permutaciones de cuatro letras que pueden formarse con las letras de la palabra *Tennessee*.

6.80 Probar que $1 - \binom{n}{1} + \binom{n}{2} - \binom{n}{3} + \cdots + (-1)^n\binom{n}{n} = 0$.

APROXIMACIÓN DE STIRLING PARA $n!$

6.81 ¿De cuántas maneras se pueden seleccionar 30 individuos de un grupo de 100 individuos?

6.82 Mostrar que para valores grandes de n $\binom{2n}{n} = 2^{2n}/\sqrt{\pi n}$, aproximadamente.

PROBLEMAS MISCELÁNEOS

6.83 De una baraja de 52 cartas se extraen tres cartas. Encontrar la probabilidad de que: *a*) dos sean sotas y una sea rey, *b*) todas sean de un mismo palo, *c*) todas sean de palos diferentes y *d*) por lo menos dos sean ases.

6.84 Encontrar la probabilidad de que de cuatro lanzamientos de un par de dados, por lo menos en dos se obtenga como suma 7.

6.85 Si 10% de los remaches que produce una máquina están defectuosos, ¿cuál es la probabilidad de que de 5 remaches tomados al azar: *a*) ninguno esté defectuoso, *b*) 1 esté defectuoso y *c*) por lo menos 2 estén defectuosos?

6.86 *a*) Dar un espacio muestral para los resultados de 2 lanzamientos de una moneda empleando 1 para representar "cara" y 0 para representar "cruz".

b) A partir de este espacio muestral, determinar la probabilidad de por lo menos una cara.

c) ¿Se puede dar el espacio muestral para los resultados de tres lanzamientos de una moneda? Determinar con ayuda de este espacio muestral la probabilidad de cuando mucho dos caras.

6.87 En una encuesta realizada con 200 votantes, se obtuvo la información siguiente acerca de tres candidatos (A, B y C) de un partido que competían por tres puestos diferentes:

28 a favor de A y B	122 a favor de B o C, pero no de A
98 a favor de A o B, pero no de C	64 a favor de C, pero no de A o B
42 a favor de B, pero no de A o C	14 a favor de A y C, pero no de B

¿Cuántos de los votantes estuvieron a favor de: *a*) los tres candidatos, *b*) A sin tener en cuenta a B o C, *c*) B sin tener en cuenta a A o C, *d*) C sin tener en cuenta a A o B, *e*) A y B, pero no de C, y *f*) sólo uno de los candidatos?

6.88 *a*) Probar que para dos eventos E_1 y E_2 cualquiera, $\Pr\{E_1 + E_2\} \leq \Pr\{E_1\} + \Pr\{E_2\}$.

b) Generalizar los resultados del inciso *a*).

6.89 Sean E_1, E_2 y E_3 tres eventos diferentes y se sabe que por lo menos uno de ellos ha ocurrido. Supóngase que cualquiera de estos eventos tiene como resultado otro evento A, que también se sabe que ya ha ocurrido. Si todas las probabilidades $\Pr\{E_1\}$, $\Pr\{E_2\}$, $\Pr\{E_3\}$ y $\Pr\{A|E_1\}$, $\Pr\{A|E_2\}$, $\Pr\{A|E_3\}$ se suponen conocidas, probar que

$$\Pr\{E_1|A\} = \frac{\Pr\{E_1\}\ \Pr\{A|E_1\}}{\Pr\{E_1\}\ \Pr\{A|E_1\} + \Pr\{E_2\}\ \Pr\{A|E_2\} + \Pr\{E_3\}\ \Pr\{A|E_3\}}$$

existiendo resultados similares para $\Pr\{E_2|A\}$ y $\Pr\{E_3|A\}$. Esto se conoce como *regla* o *teorema de Bayes*, y es útil para calcular las probabilidades de diversos E_1, E_2 y E_3 hipotéticos que han dado como resultado un evento A. Este resultado puede generalizarse.

6.90 Se tienen tres joyeros idénticos con dos cajones cada uno. En cada cajón del primer joyero hay un reloj de oro. En cada cajón del segundo joyero hay un reloj de plata. En un cajón del tercer joyero hay un reloj de oro, y en el otro cajón hay un reloj de plata. Si se toma al azar uno de los joyeros, se abre uno de los cajones y se encuentra que contiene un reloj de plata, ¿cuál es la probabilidad de que en el otro cajón se encuentre un reloj de oro? [*Sugerencia:* Emplear el problema 6.89.]

6.91 Encontrar la probabilidad de ganar en un sorteo en el que hay que elegir seis números, en cualquier orden, de entre los números 1, 2, 3, . . . , 40.

6.92 Repetir el problema 6.91 si hay que escoger: *a*) cinco, *b*) cuatro y *c*) tres números.

6.93 En un juego de póquer, a cada jugador se le dan cinco cartas de un juego de 52 naipes. Determinar las posibilidades en contra de que a un jugador le toque:

a) Una flor imperial (as, rey, reina, sota y 10 de un mismo palo).

b) Una corrida (5 cartas consecutivas y del mismo palo; por ejemplo, 3, 4, 5, 6 y 7 de espadas).

c) Un póquer (por ejemplo 4 sietes).

d) Un full (3 de un tipo y 2 de otro; por ejemplo, 3 reyes y 2 dieces).

6.94 A y B acuerdan encontrarse entre 3 y 4 de la tarde y también acuerdan que no esperarán al otro más de 10 minutos. Determinar la probabilidad de que se encuentren.

6.95 En un segmento de recta de longitud $a > 0$ se seleccionan en forma aleatoria dos puntos. Encontrar la probabilidad de que los tres segmentos de recta que se forman puedan ser los lados de un triángulo.

6.96 Un tetraedro regular consta de cuatro lados. Cada lado tiene la misma posibilidad de caer hacia abajo cuando el tetraedro es lanzado y vuelve al reposo. En cada uno de los lados hay uno de los números 1, 2, 3 o 4. Sobre una mesa se lanzan tres tetraedros regulares. Sea X la suma de las caras que caen hacia abajo. Dar la distribución de probabilidad de X.

6.97 En el problema 6.96, encontrar el valor esperado de X.

6.98 En una encuesta realizada a un grupo de personas se encontró que 25% eran fumadoras y bebedoras, 10% eran fumadoras pero no bebedoras, y 33% eran bebedoras pero no fumadoras. ¿Qué porcentaje eran fumadoras o bebedoras o ambas cosas?

6.99 Acme electronics fabrica MP3 en tres lugares. La fábrica situada en Omaha fabrica el 50% de los MP3, 1% de los cuales tienen algún defecto. La fábrica en Memphis fabrica el 30%, el 2% de éstos tienen algún defecto. La fábrica en Fort Meyers fabrica el 20% y el 3% de éstos tienen algún defecto. Si se toma al azar un MP3, ¿cuál es la probabilidad de que tenga algún defecto?

6.100 Con respecto al problema 6.99: se encuentra un MP3 que tiene algún defecto, ¿cuál es la probabilidad de que haya sido fabricado en Fort Meyers?

LAS DISTRIBUCIONES BINOMIAL, NORMAL Y DE POISSON

7

LA DISTRIBUCIÓN BINOMIAL

Si p es la probabilidad de que en un solo ensayo ocurra un evento (llamada la probabilidad de *éxito*) y $q = 1 - p$ es la probabilidad de que este evento no ocurra en un solo ensayo (llamada probabilidad de *fracaso*), entonces la probabilidad de que el evento ocurra exactamente X veces en N ensayos (es decir, que ocurran X éxitos y $N - X$ fracasos) está dada por

$$p(X) = \binom{N}{X} p^X q^{N-X} = \frac{N!}{X!\,(N-X)!}\, p^X q^{N-X} \tag{1}$$

donde $X = 0, 1, 2, \ldots, N$; $N! = N(N-1)(N-2)\cdots 1$; y $0! = 1$ por definición (ver problema 6.34).

EJEMPLO 1 La probabilidad de obtener exactamente dos caras en seis lanzamientos de una moneda es

$$\binom{6}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{6-2} = \frac{6}{2!\,4!}\left(\frac{1}{2}\right)^6 = \frac{15}{64}$$

empleando la fórmula (*1*) con $N = 6$, $X = 2$ y $p = q = \frac{1}{2}$.

Usando EXCEL, la evaluación de la probabilidad de 2 caras en 6 lanzamientos se obtiene de la siguiente manera: =BINOMDIST(2,6,0,5,0), donde la función BINOMDIST tiene 4 parámetros.

El primer parámetro es el número de éxitos, el segundo es el número de ensayos, el tercero es la probabilidad de éxito y el cuarto es 0 o 1. Cero da la probabilidad del número de éxitos y uno da la probabilidad acumulada. La función =BINOMDIST(2,6,0.5,0) da 0.234375 que es lo mismo que 15/64.

EJEMPLO 2 La probabilidad de obtener por lo menos 4 caras en 6 lanzamientos de una moneda es

$$\binom{6}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^{6-4} + \binom{6}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^{6-5} + \binom{6}{6}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^{6-6} = \frac{15}{64} + \frac{6}{64} + \frac{1}{64} = \frac{11}{32}$$

A la distribución de probabilidad discreta (*1*) suele llamársele *distribución binomial*, debido a que a $X = 0, 1, 2, \ldots, N$ le corresponden los términos sucesivos de la *fórmula binomial* o *expansión binomial*,

$$(q+p)^N = q^N + \binom{N}{1} q^{N-1}p + \binom{N}{2} q^{N-2}p^2 + \cdots + p^N \tag{2}$$

donde $1, \binom{N}{1}, \binom{N}{2}, \ldots$ se conocen como *coeficientes binomiales*.

Empleando EXCEL, la solución es =1-BINOMDIST(3,6,0.5,1) o 0.34375 que es lo mismo que 11/32. Como $\Pr\{X \geq 4\} = 1 - \Pr\{X \leq 3\}$ y BINOMDIST(3,6,0.5,1) $= \Pr\{X \leq 3\}$, este cálculo da la probabilidad de obtener por lo menos 4 caras.

EJEMPLO 3

$$\begin{aligned}(q+p)^4 &= q^4 + \binom{4}{1} q^3 p + \binom{4}{2} q^2 p^2 + \binom{4}{3} q p^3 + p^4 \\ &= q^4 + 4q^3 p + 6q^2 p^2 + 4q p^3 + p^4\end{aligned}$$

En la tabla 7.1 se enumeran algunas de las propiedades de las distribuciones binomiales.

Tabla 7.1 Distribución binomial

Media	$\mu = Np$
Varianza	$\sigma^2 = Npq$
Desviación estándar	$\sigma = \sqrt{Npq}$
Coeficiente momento de sesgo	$\alpha_3 = \dfrac{q-p}{\sqrt{Npq}}$
Coeficiente momento de curtosis	$\alpha_4 = 3 + \dfrac{1-6pq}{Npq}$

EJEMPLO 4 En 100 lanzamientos de una moneda, el número medio de caras es $\mu = Np = (100)(\frac{1}{2}) = 50$; éste es el número *esperado* de caras en 100 lanzamientos de una moneda. La desviación estándar es $\sigma = \sqrt{Npq} = \sqrt{(100)(\frac{1}{2})(\frac{1}{2})} = 5$.

LA DISTRIBUCIÓN NORMAL

Uno de los ejemplos más importantes de distribución de probabilidad continua es la *distribución normal, curva normal* o *distribución gaussiana*, que se define mediante la ecuación

$$Y = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-1/2(X-\mu)^2/\sigma^2} \tag{3}$$

donde $\mu =$ media, $\sigma =$ desviación estándar, $\pi = 3.14159\cdots$ y $e = 2.71828\cdots$. El total del área, que está limitada por la curva (*3*) y por el eje X es 1; por lo tanto, el área bajo la curva comprendida entre $X = a$ y $X = b$, donde $a < b$ representa la probabilidad de que X se encuentre entre a y b. Esta probabilidad se denota por $\Pr\{a < X < b\}$.

Si la variable X se expresa en términos de unidades estándar $[z = (X - \mu)/\sigma]$, en lugar de la ecuación (*3*) se tiene la llamada *forma estándar*:

$$Y = \frac{1}{\sqrt{2\pi}}\, e^{-1/2z^2} \tag{4}$$

En estos casos se dice que z está *distribuida normalmente y que tiene media 0 y varianza 1*. En la figura 7-1 se presenta la gráfica de esta curva normal estándar; también se muestra que las áreas comprendidas entre $z = -1$ y $z = +1$, $z = -2$ y $z = +2$, y $z = -3$ y $z = +3$ son iguales, respectivamente, a 68.27%, 95.45% y 99.73% del área total, que es 1. En la tabla que se presenta en el apéndice II se dan las áreas bajo esta curva entre $z = 0$ y cualquier valor positivo de z. Con ayuda de esta tabla se encuentra el área entre dos valores de z cualesquiera, empleando la simetría de la curva respecto a $z = 0$.

En la tabla 7.2 se enumeran algunas propiedades de la distribución normal dada por la ecuación (*3*).

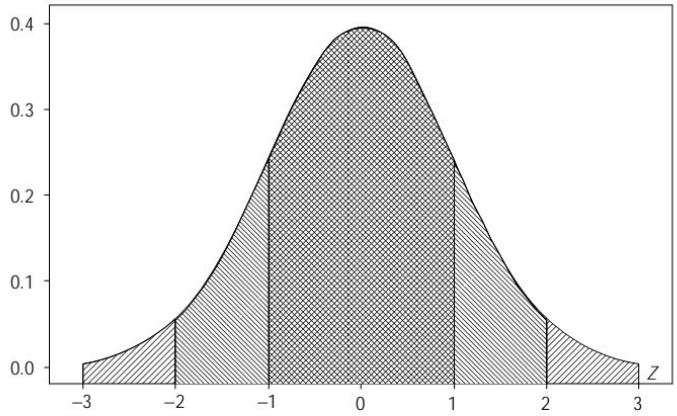


Figura 7-1 Curva normal estándar: 68.27% del área está entre $z = -1$ y $z = 1$, 95.45% del área está entre $z = -2$ y $z = 2$ y 99.73% del área está entre $z = -3$ y $z = 3$.

Tabla 7.2 Distribución normal

Media	μ
Varianza	σ^2
Desviación estándar	σ
Coeficiente momento de sesgo	$\alpha_3 = 0$
Coeficiente momento de curtosis	$\alpha_4 = 3$
Desviación media	$\sigma\sqrt{2/\pi} = 0.7979\sigma$

RELACIÓN ENTRE LAS DISTRIBUCIONES BINOMIAL Y NORMAL

Si N es grande y si ni p ni q tienen valores muy cercanos a cero, la distribución binomial puede ser aproximada por una distribución normal con la variable estandarizada dada por

$$z = \frac{X - Np}{\sqrt{Npq}}$$

A medida que crece N, la aproximación mejora y en el caso límite es exacta; esto se muestra en las tablas 7.1 y 7.2, de donde es claro que a medida que N aumenta, el sesgo y la curtosis de la distribución binomial se aproximan al sesgo y a la curtosis de la distribución normal. En la práctica, la aproximación es muy buena si tanto Np como Nq son mayores a 5.

EJEMPLO 5 En la figura 7-2 se muestra la distribución binomial correspondiente a $N = 16$ y $p = 0.5$, ilustrando las probabilidades de obtener X caras en 16 lanzamientos de una moneda, así como la distribución normal con media 8 y desviación estándar 2. Obsérvese lo semejante que son ambas distribuciones. X es binomial, con media $= Np = 16(0.5) = 8$ y desviación estándar $\sqrt{Npq} = \sqrt{16(0.5)(0.5)} = 2$. Y es una curva normal con media $= 8$ y desviación estándar 2.

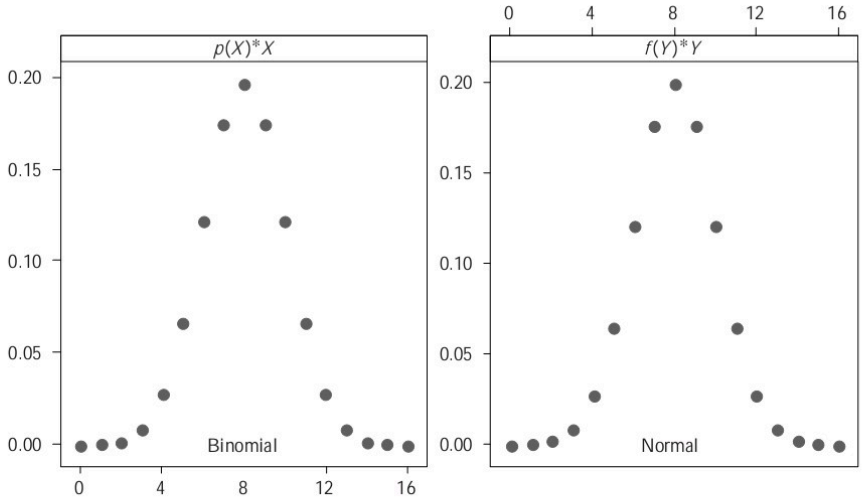


Figura 7-2 Gráfica de una curva binomial correspondiente a $N = 16$ y $p = 0.5$ y una curva normal con media = 8 y desviación estándar = 2.

LA DISTRIBUCIÓN DE POISSON

La distribución de probabilidad discreta

$$p(X) = \frac{\lambda^X e^{-\lambda}}{X!} \qquad X = 0, 1, 2, \ldots \tag{5}$$

donde $e = 2.71828\cdots$ y λ es una constante dada, se conoce como *distribución de Poisson* en honor a Siméon-Denis Poisson, quien la descubrió a comienzos del siglo XIX. Los valores de $p(X)$ pueden calcularse empleando la tabla del apéndice VIII (la cual da los valores de $e^{-\lambda}$ para diversos valores de λ) o usando logaritmos.

EJEMPLO 6 El número de personas por día que llegan a una sala de urgencias tiene una distribución de Poisson con media 5. Hallar la probabilidad de que cuando mucho lleguen tres personas por día y la probabilidad de que por lo menos lleguen 8 personas por día. La probabilidad de que cuando mucho lleguen 3 personas es $\Pr\{X \le 3\} = e^{-5}\{5^0/0! + 5^1/1! + 5^2/2! + 5^3/3!\}$. De acuerdo con el apéndice VIII, $e^{-5} = 0.006738$ y $Pr\{X \le 3\} = 0.006738\{1 + 5 + 12.5 + 20.8333\} = 0.265$. Empleando MINITAB, la secuencia "**Calc ⇒ Probability distribution ⇒ Poisson**" da la caja de diálogo de la distribución de Poisson que se llena como se muestra en la figura 7-3.

El resultado que se obtiene es el siguiente:

Función de distribución acumulada

```
Poisson with mean = 5

x     P(X<=x)
3     0.265026
```

El resultado es el mismo que el hallado usando el apéndice VIII.

La probabilidad de que lleguen por lo menos 8 personas por día es $\Pr\{X \geq 8\} = 1 - \Pr\{X \leq 7\}$. Empleando MINITAB se encuentra:

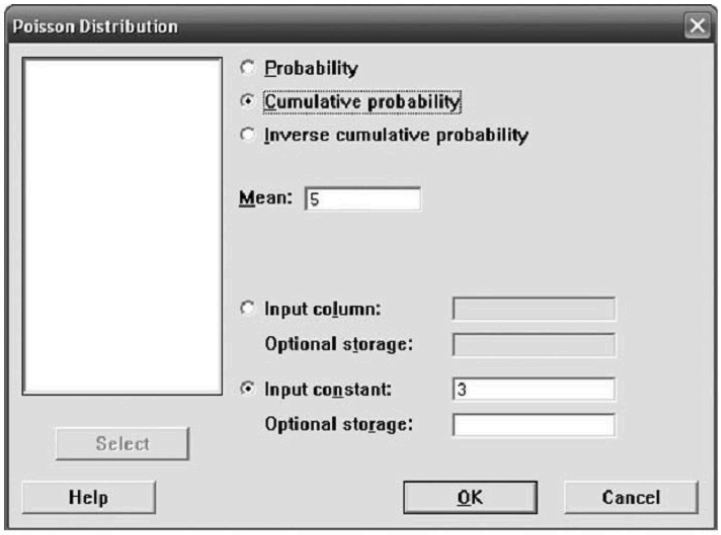


Figura 7-3 MINITAB, cuadro de diálogo para la distribución de Poisson.

Función de distribución acumulada

```
Poisson with mean = 5
x     P(X < = x)
7     0.866628
```

$\Pr\{X \geq 8\} = 1 - 0.867 = 0.133.$

En la tabla 7.3 se enumeran algunas de las propiedades de la distribución de Poisson.

Tabla 7.3 Distribución de Poisson

Media	$\mu = \lambda$
Varianza	$\sigma^2 = \lambda$
Desviación estándar	$\sigma = \sqrt{\lambda}$
Coeficiente momento de sesgo	$\alpha_3 = 1/\sqrt{\lambda}$
Coeficiente momento de curtosis	$\alpha_4 = 3 + 1/\lambda$

RELACIÓN ENTRE LAS DISTRIBUCIONES BINOMIAL Y DE POISSON

En la distribución binomial (*1*), si *N* es grande, pero la probabilidad *p* de la ocurrencia de un evento es cercana a 0, con lo que $q = 1 - p$ es cercana 1, al evento se le llama *evento raro*. En la práctica se considera que un evento es raro si el número de ensayos es por lo menos 50 ($N \geq 50$) mientras que *Np* es menor a cinco. En tales casos la distribución binomial (*1*) se aproxima con la distribución de Poisson (*5*) con $\lambda = Np$. Esto se comprueba comparando las tablas 7.1 y 7.3, ya que sustituyendo $\lambda = Np$, $q \approx 1$ y $p \approx 0$ en la tabla 7.1 se obtienen los resultados de la tabla 7.3.

Como existe una relación entre las distribuciones binomial y normal, también existe una relación entre las distribuciones de Poisson y normal. En efecto, se puede demostrar que a medida que λ aumenta indefinidamente, la distribución de Poisson se aproxima a la distribución normal con variable estandarizada $(X - \lambda)/\sqrt{\lambda}$.

LA DISTRIBUCIÓN MULTINOMIAL

Si los eventos $E_1, E_2, \ldots, E_K$ pueden ocurrir con probabilidades $p_1, p_2, \ldots, p_K$, respectivamente, entonces la probabilidad de que $E_1, E_2, \ldots, E_K$ ocurran $X_1, X_2, \ldots, X_K$ veces, respectivamente, es

$$\frac{N!}{X_1!X_2!\cdots X_K!}p_1^{X_1}p_2^{X_2}\cdots p_K^{X_K} \tag{6}$$

donde $X_1 + X_2 + \cdots + X_K = N$. A esta distribución, que es una generalización de la distribución binomial, se le llama *distribución multinomial* debido a que la ecuación (*6*) es el término general en la *expansión multinomial* $(p_1 + p_2 + \cdots + p_K)^N$.

EJEMPLO 7 Si un dado se lanza 12 veces, la probabilidad de obtener cada uno de los números 1, 2, 3, 4, 5 y 6 exactamente dos veces es

$$\frac{12!}{2!2!2!2!2!2!}\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2 = \frac{1\,925}{559\,872} = 0.00344$$

Los números *esperados* de veces que ocurrirán $E_1, E_2, \ldots, E_K$ en N ensayos son $Np_1, Np_2, \ldots, Np_K$, respectivamente.

Para obtener este resultado se puede emplear EXCEL de la manera siguiente: se usan =MULTINOMIAL(2,2,2,2,2,2) para evaluar $\frac{12!}{2!2!2!2!2!2!}$, con lo que se obtiene 7 484 400. Esto se divide después entre 6^{12}, que es 2 176 782 336. El cociente es 0.00344.

AJUSTE DE DISTRIBUCIONES DE FRECUENCIAS MUESTRALES MEDIANTE DISTRIBUCIONES TEÓRICAS

Cuando por medio de un razonamiento probabilístico, o de alguna otra manera, se tiene idea de la distribución de una población, tal distribución teórica (también llamada distribución *modelo* o *esperada*) puede ajustarse a distribuciones de frecuencias de una muestra obtenidas de una población. El método utilizado consiste, por lo general, en emplear la media y la desviación estándar de la muestra para estimar la media y la desviación estándar de la población (ver problemas 7.31, 7.33 y 7.34).

Para probar la *bondad de ajuste* de las distribuciones teóricas se usa la prueba *ji-cuadrada* (que se presenta en el capítulo 12). Cuando se quiere determinar si una distribución normal representa un buen ajuste para datos dados es conveniente emplear *papel gráfico de curva normal*, o *papel gráfico de probabilidad*, como se le suele llamar (ver problema 7.32).

PROBLEMAS RESUELTOS

LA DISTRIBUCIÓN BINOMIAL

7.1 Evaluar las expresiones siguientes:

a) $5!$ *c*) $\binom{8}{3}$ *e*) $\binom{4}{4}$

b) $\frac{6!}{2!\,4!}$ *d*) $\binom{7}{5}$ *f*) $\binom{4}{0}$

SOLUCIÓN

a) $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$

b) $\frac{6!}{2!\,4!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(2 \cdot 1)(4 \cdot 3 \cdot 2 \cdot 1)} = \frac{6 \cdot 5}{2 \cdot 1} = 15$

c) $\binom{8}{3} = \frac{8!}{3!\,(8-3)!} = \frac{8!}{3!\,5!} = \frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(3 \cdot 2 \cdot 1)(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)} = \frac{8 \cdot 7 \cdot 6}{3 \cdot 2 \cdot 1} = 56$

d) $\binom{7}{5} = \frac{7!}{5!\,2!} = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)(2 \cdot 1)} = \frac{7 \cdot 6}{2 \cdot 1} = 21$

e) $\binom{4}{4} = \frac{4!}{4!\,0!} = 1$ ya que por definición, $0! = 1$

f) $\binom{4}{0} = \frac{4!}{0!\,4!} = 1$

7.2 Supóngase que 15% de la población es zurda. Encontrar la probabilidad de que en un grupo de 50 individuos haya: *a*) cuando mucho 10 zurdos, *b*) por lo menos 5 zurdos, *c*) entre 3 y 6 zurdos y *d*) exactamente 5 zurdos. Usar EXCEL para hallar las soluciones.

SOLUCIÓN

a) La expresión de EXCEL =BINOMDIST(10,50,0.15,1) da $\Pr\{X \leq 10\}$ que es 0.8801.

b) Se pide hallar $\Pr\{X \geq 5\}$ que es igual a $1 - \Pr\{X \leq 4\}$, ya que $X \geq 5$ y $X \leq 4$ son eventos complementarios. La expresión de EXCEL para obtener el resultado buscado es =1-BINOMDIST(4,50,0.15,1), que da 0.8879.

c) Se pide hallar $\Pr\{3 \leq X \leq 6\}$ que es igual a $\Pr\{X \leq 6\} - \Pr\{X \leq 2\}$. La expresión de EXCEL para obtener el resultado buscado es =BINOMDIST(6,50,0.15,1)-BINOMDIST(2,50,0.15,1) que proporciona 0.3471.

d) La expresión de EXCEL =BINOMDIST(5,50,0.15,0) da $\Pr\{X = 5\}$ que da 0.1072.

7.3 Hallar la probabilidad de que en cinco lanzamientos de un dado aparezca un 3: *a*) ninguna vez, *b*) una vez, *c*) dos veces, *d*) tres veces, *e*) cuatro veces, *f*) cinco veces y *g*) dar la solución empleando MINITAB.

SOLUCIÓN

La probabilidad de obtener un 3 en un solo lanzamiento $= p = \frac{1}{6}$ y la probabilidad de no obtener un 3 en un solo lanzamiento $= q = 1 - p = \frac{5}{6}$; por lo tanto:

a) $\Pr\{\text{aparezca un 3 cero veces}\} = \binom{5}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^5 = (1)(1)\left(\frac{5}{6}\right)^5 = \frac{3\,125}{7\,776}$

b) $\Pr\{\text{aparezca un 3 una vez}\} = \binom{5}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^4 = (5)\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^4 = \frac{3\,125}{7\,776}$

c) $\Pr\{\text{aparezca un 3 dos veces}\} = \binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 = (10)\left(\frac{1}{36}\right)\left(\frac{125}{216}\right) = \frac{625}{3\,888}$

d) $\Pr\{\text{aparezca un 3 tres veces}\} = \binom{5}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 = (10)\left(\frac{1}{216}\right)\left(\frac{25}{36}\right) = \frac{125}{3\,888}$

e) $\Pr\{\text{aparezca un 3 cuatro veces}\} = \binom{5}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^1 = (5)\left(\frac{1}{1\,296}\right)\left(\frac{5}{6}\right) = \frac{25}{7\,776}$

f) $\Pr\{\text{aparezca un 3 cinco veces}\} = \binom{5}{5}\left(\frac{1}{6}\right)^5\left(\frac{5}{6}\right)^0 = (1)\left(\frac{1}{7\,776}\right)(1) = \frac{1}{7\,776}$

Obsérvese que estas probabilidades corresponden a los términos de la expansión binomial

$$\left(\frac{5}{6}+\frac{1}{6}\right)^5 = \left(\frac{5}{6}\right)^5 + \binom{5}{1}\left(\frac{5}{6}\right)^4\left(\frac{1}{6}\right) + \binom{5}{2}\left(\frac{5}{6}\right)^3\left(\frac{1}{6}\right)^2 + \binom{5}{3}\left(\frac{5}{6}\right)^2\left(\frac{1}{6}\right)^3 + \binom{5}{4}\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)^4 + \left(\frac{1}{6}\right)^5 = 1$$

g) En la columna C1 se ingresan los enterados 0 a 5 y después se llena el cuadro de diálogo para la distribución binomial como se indica en la figura 7-4.

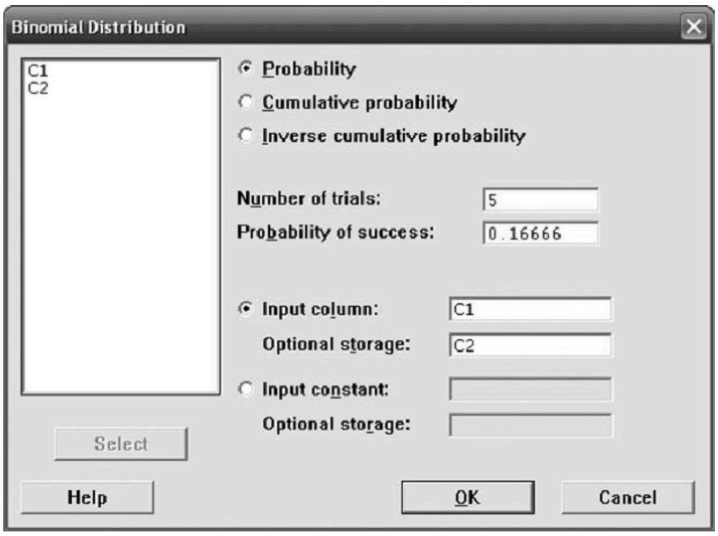


Figura 7-4 MINITAB, cuadro de diálogo para el problema 7.3*g*).

En la hoja de cálculo se obtiene el siguiente resultado:

```
C1   C2
0    0.401894
1    0.401874
2    0.160742
3    0.032147
4    0.003215
5    0.000129
```

Mostrar que las fracciones dadas en los incisos *a*) a *f*) se transforman en los decimales que se obtienen con MINITAB.

7.4 Escribir la expansión binomial de *a*) $(q+p)^4$ y de *b*) $(q+p)^6$.

SOLUCIÓN

a)
$$(q+p)^4 = q^4 + \binom{4}{1}q^3p + \binom{4}{2}q^2p^2 + \binom{4}{3}qp^3 + p^4$$
$$= q^4 + 4q^3p + 6q^2p^2 + 4qp^3 + p^4$$

b)
$$(q+p)^6 = q^6 + \binom{6}{1}q^5p + \binom{6}{2}q^4p^2 + \binom{6}{3}q^3p^3 + \binom{6}{4}q^2p^4 + \binom{6}{5}qp^5 + p^6$$
$$= q^6 + 6q^5p + 15q^4p^2 + 20q^3p^3 + 15q^2p^4 + 6qp^5 + p^6$$

Los coeficientes 1, 4, 6, 4, 1 y 1, 6, 15, 20, 15, 6, 1 son los *coeficientes binomiales* correspondientes a $N = 4$ y $N = 6$, respectivamente. Si se escriben estos coeficientes para $N = 0, 1, 2, 3, \ldots$, como se muestra en la figura siguiente, se obtiene el llamado *triángulo de Pascal*. Obsérvese que en cada renglón el primero y el último número es un 1, y que cada número se obtiene sumando los números que se encuentran a la izquierda y a la derecha en el renglón superior.

$$
\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1
\end{array}
$$

7.5 Encontrar la probabilidad de que en una familia con cuatro hijos haya: *a*) por lo menos un niño y *b*) por lo menos un niño y una niña. Supóngase que la probabilidad de que nazca un niño varón es $\frac{1}{2}$.

SOLUCIÓN

a)
$$\Pr\{1 \text{ niño}\} = \binom{4}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^3 = \frac{1}{4} \qquad \Pr\{3 \text{ niños}\} = \binom{4}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right) = \frac{1}{4}$$

$$\Pr\{2 \text{ niños}\} = \binom{4}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 = \frac{3}{8} \qquad \Pr\{4 \text{ niños}\} = \binom{4}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^0 = \frac{1}{16}$$

Por lo tanto,
$$\Pr\{\text{por lo menos 1 niño}\} = \Pr\{1 \text{ niño}\} + \Pr\{2 \text{ niños}\} + \Pr\{3 \text{ niños}\} + \Pr\{4 \text{ niños}\}$$
$$= \frac{1}{4} + \frac{3}{8} + \frac{1}{4} + \frac{1}{16} = \frac{15}{16}$$

Otro método

$$\Pr\{\text{por lo menos 1 niño}\} = 1 - \Pr\{\text{ningún niño}\} = 1 - \left(\frac{1}{2}\right)^4 = 1 - \frac{1}{16} = \frac{15}{16}$$

b) $\Pr\{\text{por lo menos 1 niño y 1 niña}\} = 1 - \Pr\{\text{ningún niño}\} - \Pr\{\text{ninguna niña}\} = 1 - \frac{1}{16} - \frac{1}{16} = \frac{7}{8}$

7.6 De 2 000 familias con cuatro hijos cada una, ¿cuántas se esperaría que tuvieran: *a*) por lo menos un niño, *b*) dos niños, *c*) 1 o 2 niñas y *d*) ninguna niña? Consultar el problema 7.5*a*).

SOLUCIÓN

a) Número esperado de familias por lo menos con 1 niño $= 2\,000(\frac{15}{16}) = 1\,875$

b) Número esperado de familias con 2 niños $= 2\,000 \cdot \Pr\{2 \text{ niños}\} = 2\,000(\frac{3}{8}) = 750$

c) $\Pr\{1 \text{ o } 2 \text{ niñas}\} = \Pr\{1 \text{ niña}\} + \Pr\{2 \text{ niñas}\} = \Pr\{1 \text{ niño}\} + \Pr\{2 \text{ niños}\} = \frac{1}{4} + \frac{3}{8} = \frac{5}{8}$. Número esperado de familias con 1 o 2 niñas $= 2\,000(\frac{5}{8}) = 1\,250$

d) Número esperado de familias sin ninguna niña $= 2\,000(\frac{1}{16}) = 125$

7.7 Si el 20% de los tornillos que se fabrican con una máquina están defectuosos, determinar la probabilidad de que de 4 tornillos elegidos al azar: *a*) 1 tornillo esté defectuoso, *b*) 0 tornillos estén defectuosos y *c*) cuando mucho 2 tornillos estén defectuosos.

SOLUCIÓN

La probabilidad de que un tornillo esté defectuoso es $p = 0.2$ y la probabilidad de que no esté defectuoso es $q = 1 - p = 0.8$.

a)
$$\Pr\{1 \text{ de 4 tornillos esté defectuoso}\} = \binom{4}{1}(0.2)^1(0.8)^3 = 0.4096$$

b) $\Pr\{\text{0 tornillos estén defectuosos}\} = \binom{4}{0}(0.2)^0(0.8)^4 = 0.4096$

c) $\Pr\{\text{2 tornillos estén defectuosos}\} = \binom{4}{2}(0.2)^2(0.8)^2 = 0.1536$

Por lo tanto

$$\begin{aligned}\Pr\{\text{cuando mucho 2 tornillos estén defectuosos}\} &= \Pr\{\text{0 tornillos estén defectuosos}\} + \Pr\{\text{1 tornillo esté defectuoso}\} \\ &\quad + \Pr\{\text{2 tornillos estén defectuosos}\} \\ &= 0.4096 + 0.4096 + 0.1536 = 0.9728\end{aligned}$$

7.8 La probabilidad de que un estudiante que entra a la universidad se titule es 0.4. Determinar la probabilidad de que de 5 estudiantes elegidos al azar: *a*) ninguno se titule, *b*) 1 se titule, *c*) por lo menos 1 se titule, *d*) todos se titulen y *e*) emplear STATISTIX para responder los incisos *a*) a *d*).

SOLUCIÓN

a) $\Pr\{\text{ninguno se titule}\} = \binom{5}{0}(0.4)^0(0.6)^5 = 0.07776$ o aproximadamente 0.08

b) $\Pr\{\text{1 se titule}\} = \binom{5}{1}(0.4)^1(0.6)^4 = 0.2592$ o aproximadamente 0.26

c) $\Pr\{\text{por lo menos 1 se titule}\} = 1 - \Pr\{\text{ninguno se titule}\} = 0.92224$ o aproximadamente 0.92

d) $\Pr\{\text{todos se titulen}\} = \binom{5}{5}(0.4)^5(0.6)^0 = 0.01024$ o aproximadamente 0.01

e) STATISTIX sólo evalúa probabilidades binomiales acumuladas. Con el cuadro de diálogo de la figura 7-5 se obtiene la distribución de probabilidad binomial acumulada para $N = 5$, $p = 0.4$, $q = 0.6$ y $x = 0, 1, 4$ y 5.

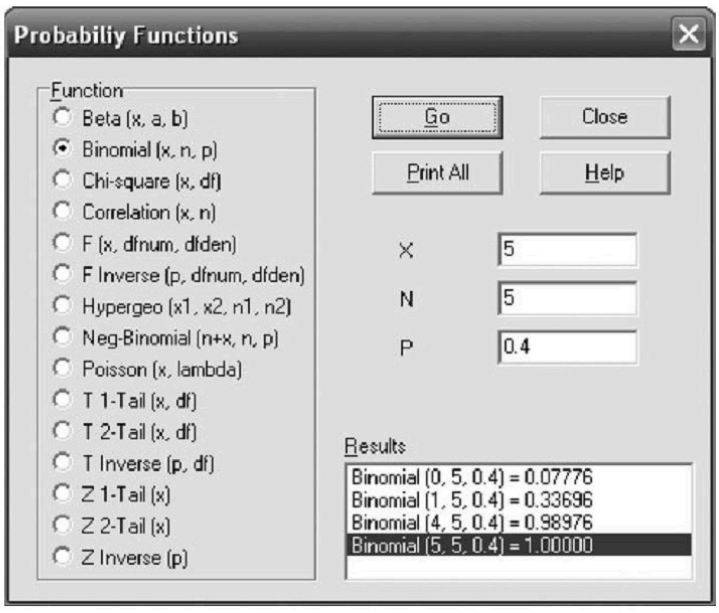


Figura 7-5 STATISTIX, cuadro de diálogo para el problema 7.8*e*).

Mediante la información obtenida en el último cuadro de diálogo, se tiene que: la probabilidad de que ninguno se titule es $\Pr\{X=0\} = \text{Binomial}(0,5,0.4) = 0.07776$. La probabilidad de que 1 se titule es $\Pr\{X=1\} = \Pr\{X\leq 1\} - \Pr\{X\leq 0\} = \text{Binomial}(1,5,0.4) - \text{Binomial}(0,5,0.4) = 0.33696 - 0.07776 = 0.2592$. La probabilidad de que por lo menos 1 se titule es $\Pr\{X\geq 1\} = 1 - \Pr\{X=0\} = 1 - \text{Binomial}(0,5,0.4) = 0.92224$. La probabilidad de que todos se titulen es $\Pr\{X=5\} = \Pr\{X\leq 5\} - \Pr\{X\leq 4\} = \text{Binomial}(5,5,0.4) - \text{Binomial}(4,5,0.4) = 1.00000 - 0.98976 = 0.01024$. Obsérvese que STATISTIX únicamente da la probabilidad binomial acumulada y también algunas de las tablas que aparecen en los libros de texto dan únicamente probabilidades binomiales acumuladas.

7.9 ¿Cuál es la probabilidad de que en 6 lanzamientos de un par de dados se obtenga como suma 9: *a*) dos veces y *b*) por lo menos 2 veces?

SOLUCIÓN

Cada una de las 6 maneras en que puede caer el primer dado se asocia con cada una de las 6 maneras en que puede caer el segundo dado; por lo tanto, hay $6 \cdot 6 = 36$ maneras en que pueden caer los dos dados. Se puede tener: 1 en el primer dado y 1 en el segundo dado, 1 en el primer dado y 2 en el segundo dado, etc., lo que se denota (1, 1), (1, 2), etcétera.

De estas 36 maneras (todas igualmente probables), la suma 9 se obtiene en 4 casos: (3, 6), (4, 5), (5, 4) y (6, 3). Por lo tanto, la probabilidad de que en un lanzamiento de los dos dados la suma sea 9 es $p = \frac{4}{36} = \frac{1}{9}$ y la probabilidad de que en un lanzamiento la suma de los dos dados no sea 9 es $q = 1 - p = \frac{8}{9}$.

a) $\Pr\{\text{2 nueves en 6 lanzamientos}\} = \binom{6}{2}\left(\frac{1}{9}\right)^2\left(\frac{8}{9}\right)^{6-2} = \frac{61\,440}{531\,441}$

b) $\Pr\{\text{por lo menos 2 nueves}\} = \Pr\{\text{2 nueves}\} + \Pr\{\text{3 nueves}\} + \Pr\{\text{4 nueves}\} + \Pr\{\text{5 nueves}\} + \Pr\{\text{6 nueves}\}$

$$= \binom{6}{2}\left(\frac{1}{9}\right)^2\left(\frac{8}{9}\right)^4 + \binom{6}{3}\left(\frac{1}{9}\right)^3\left(\frac{8}{9}\right)^3 + \binom{6}{4}\left(\frac{1}{9}\right)^4\left(\frac{8}{9}\right)^2 + \binom{6}{5}\left(\frac{1}{9}\right)^5\left(\frac{8}{9}\right)^1 + \binom{6}{6}\left(\frac{1}{9}\right)^6\left(\frac{8}{9}\right)^0$$

$$= \frac{61\,440}{531\,441} + \frac{10\,240}{531\,441} + \frac{960}{53\,1441} + \frac{48}{531\,441} = \frac{1}{531\,441} = \frac{72\,689}{531\,441}$$

Otro método

$$\Pr\{\text{por lo menos 2 nueves}\} = 1 - \Pr\{0 \text{ nueves}\} - \Pr\{1 \text{ nueve}\}$$

$$= 1 - \binom{6}{0}\left(\frac{1}{9}\right)^0\left(\frac{8}{9}\right)^6 - \binom{6}{1}\left(\frac{1}{9}\right)^1\left(\frac{8}{9}\right)^5 = \frac{72\,689}{531\,441}$$

7.10 Evaluar: *a*) $\sum_{X=0}^{N} Xp(X)$ y *b*) $\sum_{X=0}^{N} X^2 p(X)$, donde $p(X) = \binom{N}{X} p^X q^{N-X}$.

SOLUCIÓN

a) Como $q + p = 1$,

$$\sum_{X=0}^{N} Xp(X) = \sum_{X=1}^{N} X\,\frac{N!}{X!\,(N-X)!}\,p^X q^{N-X} = Np\sum_{X=1}^{N}\frac{(N-1)!}{(X-1)!(N-X)!}\,p^{X-1}q^{N-X}$$

$$= Np(q+p)^{N-1} = Np$$

b) $$\sum_{X=0}^{N} X^2 p(X) = \sum_{X=1}^{N} X^2\,\frac{N!}{X!(N-X)!}p^X q^{N-X} = \sum_{X=1}^{N}[X(X-1)+X]\,\frac{N!}{X!(N-X)!}\,p^X q^{N-X}$$

$$= \sum_{X=2}^{N} X(X-1)\,\frac{N!}{X!(N-X)!}\,p^X q^{N-X} + \sum_{X=1}^{N} X\,\frac{N!}{X!(N-X)!}\,p^X q^{N-X}$$

$$= N(N-1)p^2\sum_{X=2}^{N}\frac{(N-2)!}{(X-2)!(N-X)!}\,p^{X-2}q^{N-X} + Np = N(N-1)p^2(q+p)^{N-2} + Np$$

$$= N(N-1)p^2 + Np$$

Nota: Los resultados de los incisos *a*) y *b*) son las *esperanzas* de X y de X^2, que se denotan $E(X)$ y $E(X^2)$, respectivamente (ver capítulo 6).

7.11 Si la variable está distribuida binomialmente, determinar: *a*) su media μ y *b*) su varianza σ^2.

SOLUCIÓN

a) De acuerdo con el problema 7.10*a*),

$$\mu = \text{esperanza de la variable} = \sum_{X=0}^{N} Xp(X) = Np$$

b) Empleando $\mu = Np$ y los resultados del problema 7.10,

$$\sigma^2 = \sum_{X=0}^{N}(X-\mu)^2 p(X) = \sum_{X=0}^{N}(X^2 - 2\mu X + \mu^2)p(X) = \sum_{X=0}^{N} X^2 p(X) - 2\mu \sum_{X=0}^{N} Xp(X) + \mu^2 \sum_{X=0}^{N} p(X)$$

$$= N(N-1)p^2 + Np - 2(Np)(Np) + (Np)^2(1) = Np - Np^2 = Np(1-p) = Npq$$

Se sigue que la desviación estándar de una variable distribuida en forma binomial es $\sigma = \sqrt{Npq}$.

Otro método

De acuerdo con el problema 6.62*b*),

$$E[(X - \bar{X})]^2 = E(X^2) - [E(X)]^2 = N(N-1)p^2 + Np - N^2p^2 = Np - Np^2 = Npq$$

7.12 Si la probabilidad de que un tornillo esté defectuoso es 0.1, encontrar: *a*) la media y *b*) la desviación estándar de la distribución de los tornillos defectuosos en un total de 400 tornillos.

SOLUCIÓN

a) La media es $Np = 400(0.1) = 40$; es decir, se puede *esperar* que haya 40 tornillos defectuosos.

b) La varianza es $Npq = 400(0.1)(0.9) = 36$. Por lo tanto, la desviación estándar es $\sqrt{36} = 6$.

7.13 Encontrar el coeficiente momento de: *a*) sesgo y *b*) curtosis, de la distribución del problema 7.12.

SOLUCIÓN

a) $$\text{Coeficiente momento de sesgo} = \frac{q-p}{\sqrt{Npq}} = \frac{0.9 - 0.1}{6} = 0.133$$

Como este coeficiente es positivo, la distribución es sesgada a la derecha.

b) $$\text{Coeficiente momento de curtosis} = 3 + \frac{1 - 6pq}{Npq} = 3 + \frac{1 - 6(0.1)(0.9)}{36} = 3.01$$

Esta distribución es ligeramente *leptocúrtica* con respecto a la distribución normal (es decir, ligeramente más puntiaguda; ver capítulo 5).

LA DISTRIBUCIÓN NORMAL

7.14 En un examen final de matemáticas la media fue 72 y la desviación estándar fue 15. Determinar las puntuaciones estándar (es decir, las calificaciones en unidades de desviaciones estándar) de los estudiantes que obtuvieron: *a*) 60, *b*) 93 y *c*) 72 puntos.

SOLUCIÓN

a) $z = \dfrac{X - \bar{X}}{s} = \dfrac{60 - 72}{15} = -0.8$

b) $z = \dfrac{X - \bar{X}}{s} = \dfrac{93 - 72}{15} = 1.4$

c) $z = \dfrac{X - \bar{X}}{s} = \dfrac{72 - 72}{15} = 0$

7.15 Con los datos del problema 7.14, encontrar las calificaciones que corresponden a las siguientes puntuaciones estándar: *a*) -1 y *b*) 1.6.

SOLUCIÓN

a) $X = \bar{X} + zs = 72 + (-1)(15) = 57$ *b*) $X = \bar{X} + zs = 72 + (1.6)(15) = 96$

7.16 Supóngase que la cantidad de juegos en que participan los beisbolistas de la liga mayor durante su carrera se distribuye normalmente con media 1 500 juegos y desviación estándar 350 juegos. Emplear EXCEL para responder las preguntas siguientes. *a*) ¿Qué porcentaje participa en menos de 750 juegos? *b*) Qué porcentaje participa en más de 2 000 juegos? y *c*) Encontrar el percentil 90 de la cantidad de juegos en los que participan durante su carrera.

SOLUCIÓN

a) La expresión de EXCEL =NORMDIST(750,1 500, 350, 1) busca el área a la izquierda de 750 en una curva normal con media igual a 1 500 y desviación estándar igual a 350. La respuesta es $\Pr\{X < 750\} = 0.0161$ o bien 1.61% participa en menos de 750 juegos.

b) La expresión de EXCEL =1-NORMDIST(2 000,1 500, 350, 1) busca el área a la derecha de 2 000 en una curva normal con media igual a 1 500 y desviación estándar igual a 350. La respuesta es $\Pr\{X > 2\,000\} = 0.0766$ o bien 7.66% participa en más de 2 000 juegos.

c) La expresión de EXCEL =NORMINV(0.9,1 500, 350) busca en el eje horizontal el valor tal que a su izquierda se encuentra 90% del área bajo la curva normal con media 1 500 y desviación estándar 350. Empleando la notación del capítulo 3, $P_{90} = 1\,948.5$.

7.17 Encontrar el área bajo la curva normal en los casos siguientes.

a) Entre $z = 0.81$ y $z = 1.94$.

b) A la derecha de $z = -1.28$.

c) A la derecha de $z = 2.05$ o a la izquierda de $z = -1.44$.

Para resolver los incisos *a*) a *c*) emplear el apéndice II y EXCEL [ver la figura 7-6, incisos *a*), *b*) y *c*)].

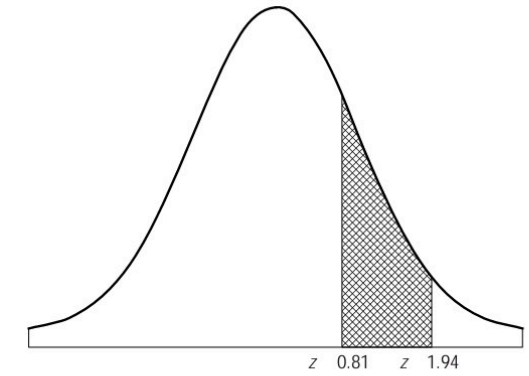
SOLUCIÓN

a) En el apéndice II, bajar por la columna z hasta llegar a 1.9; después avanzar a la derecha hasta la columna marcada con 4. El resultado 0.4738 es $\Pr\{0 \le z \le 1.94\}$. A continuación bajar por la columna z hasta llegar a 0.8; después avanzar a la derecha hasta la columna marcada con 1. El resultado 0.2910 es $\Pr\{0 \le z \le 0.81\}$. El área correspondiente a $\Pr\{0.81 \le z \le 1.94\}$ es la diferencia de ambos, $\Pr\{0 \le z \le 1.94\} - \Pr\{0 \le z \le 0.81\} = 0.4738 - 0.2910 = 0.1828$. Empleando EXCEL, la respuesta se obtiene con =NORMSDIST(1.94)-NORMSDIST(0.81) = 0.1828. Empleando EXCEL, el área $\Pr\{0.81 \le z \le 1.94\}$ es la diferencia $\Pr\{-\text{infinito} \le z \le 1.94\} - \Pr\{-\text{infinito} \le z \le 0.81\}$. Obsérvese que la tabla del apéndice II da áreas desde 0 hasta un valor positivo de z, en tanto que EXCEL da áreas desde –infinito hasta el mismo valor de z.

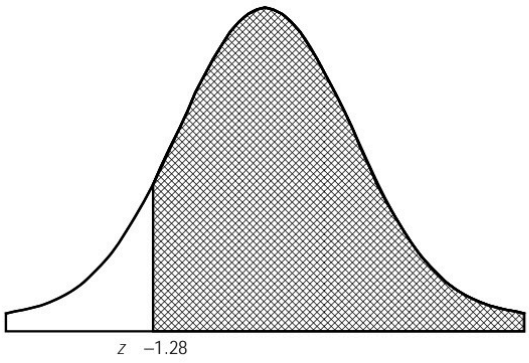
b) El área a la derecha de $z = -1.28$ es la misma área que a la izquierda de $z = 1.28$. Empleando el apéndice II, el área a la izquierda de $z = 1.28$ es $\Pr\{z \le 0\} + \Pr\{0 \le z \le 1.28\}$ o bien $0.5 + 0.3997 = 0.8997$. Usando EXCEL, $\Pr\{z \ge -1.28\} = \Pr\{z \le 1.28\}$ y $\Pr\{z \le 1.28\}$ se obtiene mediante =NORMSDIST(1.28), que da 0.8997.

c) Empleando el apéndice II, el área a la derecha de 2.05 es $0.5 - \Pr\{z \le 2.05\}$ o bien 0.5 – 0.4798 = 0.0202. El área a la izquierda de -1.44 es la misma que el área a la derecha de 1.44. El área a la derecha de 1.44 es $0.5 - \Pr\{z \le 1.44\}$ = 0.5 – 0.4251 = 0.0749. La suma de estas dos áreas en las colas es $0.0202 + 0.0749 = 0.0951$. Usando EXCEL, esta área se obtiene como sigue: =NORMSDIST(−1.44) + (1 − NORMSDIST(2.05)), que da 0.0951.

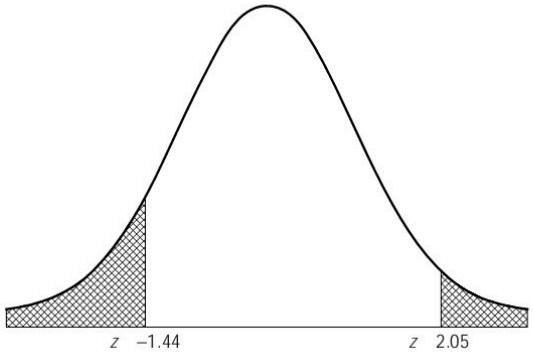
Obsérvese que en EXCEL =NORMSDIST(z) da el área a la izquierda de z bajo la curva normal estándar, en tanto que =NORMDIST(z, μ, σ, 1) da el área a la izquierda de z bajo la curva normal cuya media es μ y cuya desviación estándar de σ.



a)



b)



c)

Figura 7-6 Áreas bajo la curva normal estándar. *a*) Área entre $z = 0.81$ y $z = 1.94$; *b*) área a la derecha de $z = -1.28$; *c*) área a la izquierda de $z = -1.44$ más área a la derecha de $z = 2.05$.

7.18 La cantidad de horas, por semana, que los estudiantes de educación media ven televisión tiene una distribución normal cuya media es 20.5 horas y cuya desviación estándar es 5.5 horas. Emplear MINITAB para hallar el porcentaje que ve televisión menos de 25 horas por semana. Usar MINITAB para hallar el porcentaje que ve televisión más de 30 horas por semana. Trazar una curva que represente estos dos grupos.

SOLUCIÓN

La solución se ilustra en la figura 7-7.

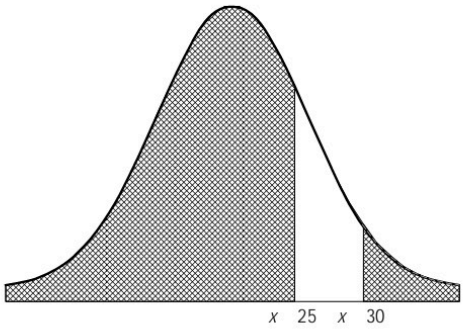


Figura 7-7 MINITAB, gráfica que muestra el grupo que ve televisión menos de 25 horas por semana y el grupo que ve televisión más de 30 horas por semana.

La secuencia "**Calc** ⇒ **Probability distributions** ⇒ **Normal**" abre el cuadro de diálogo de la distribución normal que se presenta en la figura 7-8.

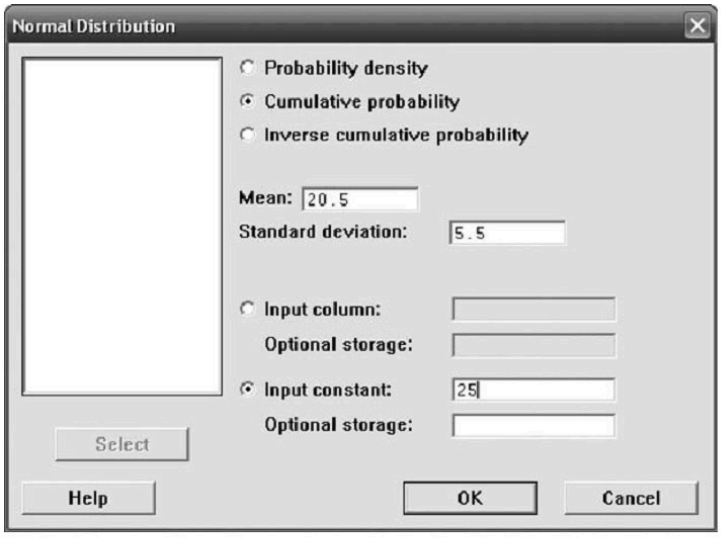


Figura 7-8 MINITAB, cuadro de diálogo para la distribución normal.

Llenando el cuadro de diálogo como se muestra en la figura 7-8 y ejecutándolo, se obtiene el siguiente resultado:

Función de distribución acumulada

```
Normal con media = 20.5 y desviación estándar = 5.5

x     P(X<=x)
25    0.793373
```

79.3% de los estudiantes de educación media ven 25 horas o menos de televisión por semana.
Si se ingresa 30 como la constante de entrada, se obtiene el resultado siguiente:

Función de distribución acumulada

```
Normal con media = 20.5 y desviación estándar = 5.5

x     P(X<=x)
30    0.957941
```

El porcentaje que ve más de 30 horas de televisión por semana es $1 - 0.958 = 0.042$ o 4.2%.

7.19 Hallar la ordenada correspondiente a la curva normal en: *a*) $z = 0.84$, *b*) $z = -1.27$ y *c*) $z = -0.05$.

SOLUCIÓN

a) En el apéndice I, bajar por la columna que tiene como encabezado z hasta llegar a la entrada 0.8; después avanzar hacia la derecha hasta la columna que tiene como encabezado 4. La entrada 0.2803 es la ordenada buscada.

b) Por simetría: (ordenada correspondiente a $z = -1.27$) = (ordenada correspondiente a $z = 1.27$) = 0.1781.

c) (La ordenada correspondiente $z = -0.05$) = (la ordenada correspondiente a $z = 0.05$) = 0.3984.

7.20 Emplear EXCEL para evaluar algunas ordenadas correspondientes a la curva normal cuya media es 13.5 y cuya desviación estándar es 2.5. Después, empleando el asistente para gráficos, graficar los puntos obtenidos. Estas gráficas representan la distribución normal de la variable X, donde X representa las horas, por semana, que los estudiantes universitarios pasan en Internet.

SOLUCIÓN

Las abscisas elegidas que van desde 6 hasta 21, a intervalos de 0.5, se ingresan en la hoja de cálculo de EXCEL en las celdas A1:A31. En la celda B1 se ingresa la expresión =NORMDIST(A1,13.5,3.5,0), se hace clic y se arrastra. Estos puntos son de la curva normal cuya media es 13.5 y cuya desviación estándar es 3.5:

6	0.001773
6.5	0.003166
7	0.005433
7.5	0.008958
8	0.01419
8.5	0.021596
9	0.03158
9.5	0.044368
10	0.059891
10.5	0.077674
11	0.096788
11.5	0.115877
12	0.13329
12.5	0.147308
13	0.156417
13.5	0.159577

14	0.156417
14.5	0.147308
15	0.13329
15.5	0.115877
16	0.096788
16.5	0.077674
17	0.059891
17.5	0.044368
18	0.03158
18.5	0.021596
19	0.01419
19.5	0.008958
20	0.005433
20.5	0.003166
21	0.001773

Para graficar estos puntos se emplea el asistente para gráficos. El resultado se muestra en la figura 7-9.

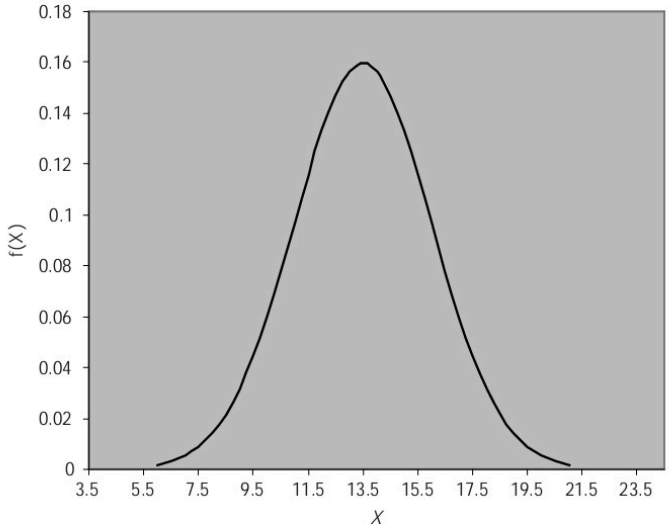


Figura 7-9 EXCEL, gráfica de la curva normal cuya media es = 13.5 y cuya desviación estándar es = 2.5.

7.21 Determinar el segundo cuartil (Q_2), el tercer cuartil (Q_3) y el percentil 90 (P_{90}) de las horas, por semana, que los estudiantes universitarios pasan en Internet. Emplear la distribución normal dada en el problema 7.20.

SOLUCIÓN

El segundo cuartil o percentil 50 de una distribución normal corresponde al centro de la curva. Debido a la simetría de esta distribución, coincide con el punto en el que se encuentra la media. En el caso del uso de Internet, éste será 13.5 horas por semana. Para hallar el percentil 50, se usa la función de EXCEL =NORMINV(0.5,13.5,2.5). El percentil 50 significa que 0.5 del área se encuentra a la izquierda del segundo cuartil y que la media es 13.5 y la desviación estándar es 2.5. El resultado que da EXCEL es 13.5. Usando MINITAB se obtiene:

Función de distribución acumulada inversa

```
Normal con media = 13.5 y desviación estándar = 2.5

p(X<=x)     x
0.5         13.5
```

Esto se ilustra en la figura 7-10 *a*). En la figura 7-10*b*) se nota que el 75% del área bajo la curva está a la izquierda de Q_3. Empleando EXCEL, la función =NORMINV(0.75,13.5,2.5) da $Q_3 = 15.19$.

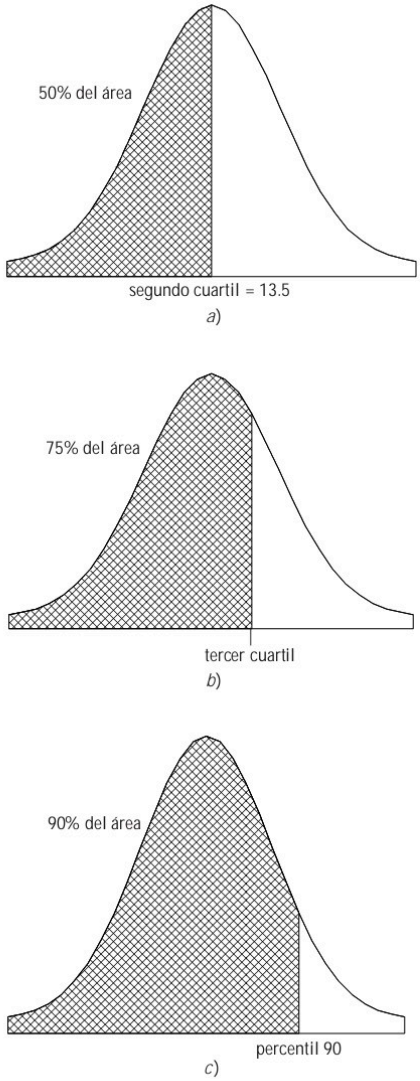


Figura 7-10 Se hallan percentiles y cuartiles usando MINITAB y EXCEL. *a*) Q_2 es el valor tal que 50% de los tiempos son menores que ese valor; *b*) Q_3 es el valor tal que 75% de los tiempos son menores que ese valor; *c*) P_{90} es el valor tal que 90% de los tiempos son menores que ese valor.

En la figura 7-10*c*) se muestra que 90% del área está a la izquierda de P_{90}. La función de EXCEL =NORMINV(0.90,13.5,2.5) da $P_{90} = 16.70$.

7.22 Usando el apéndice II, encontrar Q_3 para los datos del problema 7.21.

SOLUCIÓN

Si no se dispone de un software como EXCEL o MINITAB, es necesario trabajar con las tablas de la distribución normal estándar como único recurso.

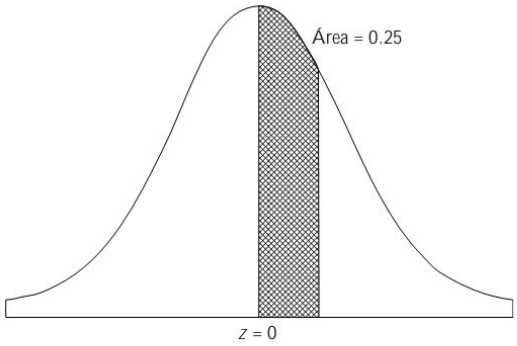


Figura 7-11 Curva normal estándar.

Usando el apéndice II en forma inversa, se ve que el área que va desde $z = 0$ hasta $z = 0.67$ es 0.2486, y el área que va desde $z = 0$ hasta $z = 0.68$ es 0.2518 (ver figura 7-11). El área desde −infinito hasta $z = 0.675$ es aproximadamente 0.75, ya que el área desde −infinito a 0 es 0.5, y el área desde 0 hasta $z = 0.675$ es 0.25. Por lo tanto, en la curva normal estándar el tercer cuartil es aproximadamente 0.675. Sea Q_3 el tercer cuartil en la curva normal cuya media es 13.5 y cuya desviación estándar es 2.5 [ver figura 7-10*b*)]. Cuando Q_3 se transforma en un valor z, se tiene $0.675 = (Q_3 - 13.5)/2.5$. Despejando en esta ecuación Q_3, se tiene $Q_3 = 2.5(0.675) + 13.5 = 15.19$, que es la misma respuesta que se obtuvo con EXCEL en el problema 7.21.

7.23 Se producen arandelas cuyo diámetro interno está distribuido normalmente con media 0.500 pulgadas (in) y desviación estándar 0.005 in. Las arandelas se consideran defectuosas si su diámetro interno es de menos de 0.490 in o si es de más de 0.510 in. Empleando tanto el apéndice II como EXCEL, hallar el porcentaje de arandelas defectuosas.

SOLUCIÓN

$$0.490 \text{ en unidades estándar es } \frac{0.490 - 0.500}{0.005} = -2.00$$

$$0.510 \text{ en unidades estándar es } \frac{0.510 - 0.500}{0.005} = 2.00$$

De acuerdo con el apéndice II, el área a la derecha de $Z = 2.00$ es $0.5 - 0.4772$, o bien 0.0228. El área a la izquierda de $Z = -2.00$ es 0.0228. El porcentaje de defectuosos es $(0.0228 + 0.0228) \times 100 = 4.56\%$. Para hallar áreas bajo una curva normal empleando el apéndice II hay que convertir los datos a la curva normal estándar para encontrar las respuestas.

Empleando EXCEL, la respuesta es = 2*NORMDIST(0.490, 0.500, 0.005, 1), que da también 4.56%.

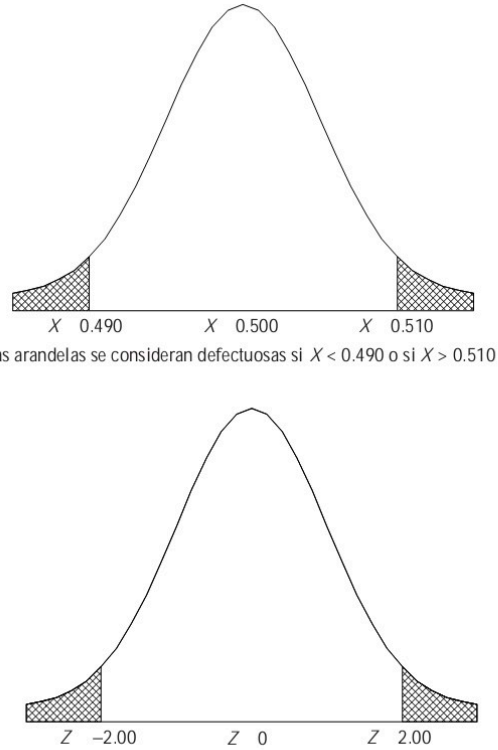


Figura 7-12 El área a la derecha de $X = 0.510$ es igual al área a la derecha de $Z = 2.000$ y el área a la izquierda de $X = 0.490$ es igual al área a la izquierda de $Z = -2.00$.

APROXIMACIÓN NORMAL A LA DISTRIBUCIÓN BINOMIAL

7.24 Empleando: *a*) la distribución binomial y *b*) la aproximación normal a la distribución binomial, encontrar la probabilidad de que en 10 lanzamientos de una moneda se obtengan 3 a 6 caras inclusive.

SOLUCIÓN

a) $$\Pr\{3 \text{ caras}\} = \binom{10}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^7 = \frac{15}{128} \qquad \Pr\{5 \text{ caras}\} = \binom{10}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^5 = \frac{63}{256}$$

$$\Pr\{4 \text{ caras}\} = \binom{10}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^6 = \frac{105}{512} \qquad \Pr\{6 \text{ caras}\} = \binom{10}{6}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^4 = \frac{105}{512}$$

Por lo tanto

$$\{\text{de entre 3 a 6 caras inclusive}\} = \frac{15}{128} + \frac{105}{512} + \frac{63}{256} + \frac{105}{512} + \frac{99}{128} = 0.7734$$

b) En la figura 7-13 se presenta la gráfica que se obtiene con EXCEL para la distribución binomial con $N = 10$ lanzamientos de una moneda.

Obsérvese que aunque la distribución binomial es una distribución discreta, esta gráfica tiene la forma de una distribución normal, que es continua. Para aproximar la probabilidad binomial de 3, 4, 5 y 6 caras mediante el área bajo la curva normal, se busca el área bajo la curva normal desde $X = 2.5$ hasta $X = 6.5$. El 0.5 que se agrega a cada lado de $X = 3$ y $X = 6$ se le llama *corrección por continuidad.* A continuación se dan los pasos a seguir para aproximar la distribución binomial mediante la distribución normal. Se elige la curva normal con media $Np = 10(0.5) = 5$ y desviación estándar $= \sqrt{Npq} = \sqrt{10(0.5)(0.5)} = 1.58$. De esta manera se elige la curva normal que tiene el mismo centro y la misma variación que la distribución binomial. Después se busca el área bajo la curva desde 2.5 hasta 6.5, como se muestra en la figura 7-14. Ésta es la aproximación normal a la distribución binomial.

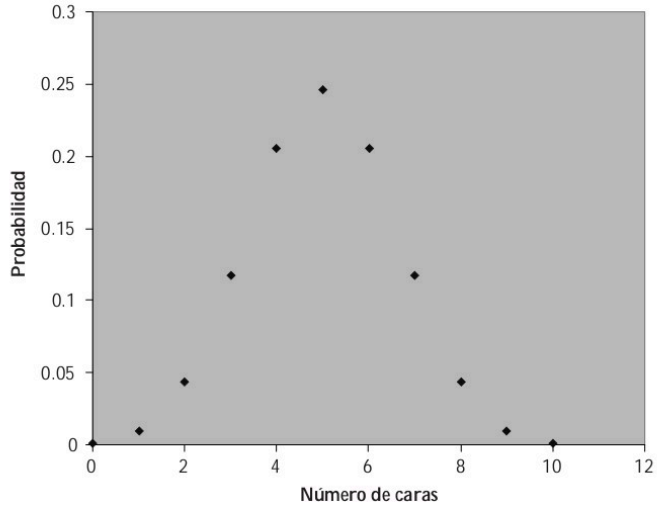


Figura 7-13 EXCEL, gráfica de la distribución binomial con $N = 10$ y $p = 0.5$.

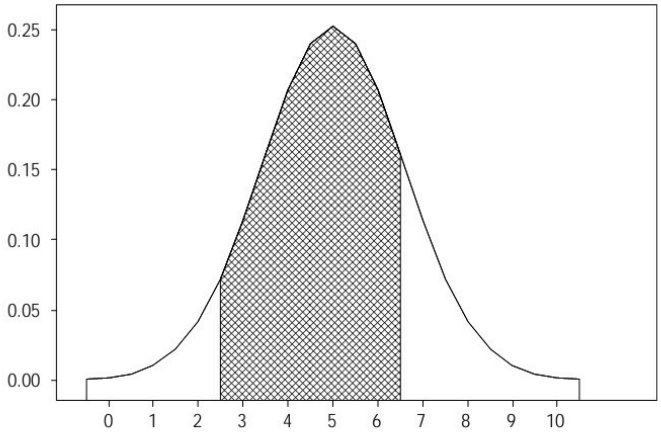


Figura 7-14 Aproximación normal para 3, 4, 5 o 6 caras cuando se lanza una moneda 10 veces.

Usando una hoja de cálculo de EXCEL, la solución se obtiene empleando =NORMDIST(6.5, 5, 1.58,1) − NORMDIST(2.5, 5, 1.58, 1), con lo que se obtiene 0.7720.

Empleando la técnica del apéndice II, los valores normales 6.5 y 2.5 se convierten primero a valores normales estándar. (En unidades estándar 2.5 es −1.58, y 6.5 en unidades estándar es 0.95.) De acuerdo con el apéndice II el área entre −1.58 y 0.95 es 0.7718. Cualquiera que sea el método que se use, el resultado es muy semejante al obtenido con la distribución binomial, 0.7734.

7.25 Se lanza una moneda 500 veces. Hallar la probabilidad de que el número de caras no sea diferente de 250: *a*) en más de 10 y *b*) en más de 30.

SOLUCIÓN

$$\mu = Np = (500)(\tfrac{1}{2}) = 250 \qquad \sigma = \sqrt{Npq} = \sqrt{(500)(\tfrac{1}{2})(\tfrac{1}{2})} = 11.18$$

a) Se busca la probabilidad de que la cantidad de caras esté entre 240 y 260 o, considerando los datos como datos continuos, entre 239.5 y 260.5. Como 239.5 en unidades estándar es $(239.5 - 250)/11.18 = -0.94$ y 260.5 en unidades estándar es 0.94, se tiene

$$\begin{aligned}\text{Probabilidad buscada} &= (\text{área bajo la curva normal entre } z = -0.94 \text{ y } z = 0.94)\\ &= (\text{dos veces el área entre } z = 0 \text{ y } z = 0.94) = 2(0.3264) = 0.6528\end{aligned}$$

b) Se busca la probabilidad de que la cantidad de caras esté entre 220 y 280 o, considerando los datos como datos continuos, entre 219.5 y 280.5. Como 219.5 en unidades estándar es $(219.5 - 250)/11.18 = -2.73$ y 280.5 en unidades estándar es 2.73, se tiene

$$\begin{aligned}\text{Probabilidad buscada} &= (\text{dos veces el área entre } z = 0 \text{ y } z = -2.73)\\ &= 2(0.4968) = 0.9936\end{aligned}$$

Por lo tanto, se puede confiar en que el número de caras no diferirá de lo esperado (250) en más de 30. De manera que si resulta que el número de caras que *realmente* se encuentra es 280, habrá razón para creer que la moneda está cargada.

7.26 Supóngase que en el grupo de edad de 1 a 4 años, el 75% usa el cinturón de seguridad de manera habitual. Hallar la probabilidad de que si se detienen, al azar, algunos automóviles que transporten pasajeros de 1 a 4 años, 70 o menos estén usando el cinturón de seguridad. Dar la solución empleando la distribución binomial así como la aproximación normal a la distribución binomial. Usar MINITAB para hallar la solución.

SOLUCIÓN

El resultado de MINITAB dado adelante muestra que la probabilidad de que 70 o menos estén usando el cinturón de seguridad es igual a 0.1495.

```
MTB > cdf 70;
SUBC> binomial 100.75.
```

Función de distribución acumulada

```
Binomial con n = 100 y p = 0.750000
    x      P( X ⇐ x)
70.00       0.1495
```

Empleando la aproximación normal a la distribución binomial, la solución se encuentra como sigue: la media de la distribución binomial es $\mu = Np = 100(0.75) = 75$ y la desviación estándar es $\sigma = \sqrt{Npq} = \sqrt{100(0.75)(0.25)} = 4.33$. El resultado de MINITAB dado adelante muestra que la aproximación normal es igual a 0.1493. Esta aproximación es muy semejante al verdadero valor.

```
MTB > cdf 70.5;
SUBC> media normal = 75 sd = 4.33
```

Función de distribución acumulada

```
Normal con media = 75.0000 y desviación estándar = 4.33000
      x      P( X ⇐ x)
70.5000       0.1493
```

LA DISTRIBUCIÓN DE POISSON

7.27 De las herramientas que se producen con determinado proceso de fabricación, 10% resultan defectuosas. Empleando: *a*) la distribución binomial y *b*) la aproximación de Poisson a la distribución binomial, hallar la probabilidad de que en una muestra de 10 herramientas elegidas al azar exactamente 2 estén defectuosas.

SOLUCIÓN

La probabilidad de que una herramienta esté defectuosa es $p = 0.1$.

a) $\Pr\{2 \text{ de } 10 \text{ herramientas defectuosas}\} = \binom{10}{2}(0.1)^2(0.9)^8 = 0.1937$ o bien 0.19

b) Con $\lambda = Np = 10(0.1) = 1$ y usando $e = 2.718$,

$$\Pr\{2 \text{ de } 10 \text{ herramientas defectuosas}\} = \frac{\lambda^X e^{-\lambda}}{X!} = \frac{(1)^2 e^{-1}}{2!} = \frac{e^{-1}}{2} = \frac{1}{2e} = 0.1839$$ o bien 0.18

En general, esta aproximación es buena si $p \leq 0.1$ y $\lambda = Np \leq 5$.

7.28 Si la probabilidad de que un individuo tenga una reacción adversa por la inyección de determinado suero es 0.001, determinar la probabilidad de que de 2 000 individuos: *a*) exactamente 3 y *b*) más de 2, sufran una reacción adversa. Usar MINITAB y hallar la respuesta empleando tanto Poisson como distribuciones binomiales.

SOLUCIÓN

a) En el siguiente resultado de MINITAB se da primero la probabilidad binomial de que exactamente 3 individuos tengan una reacción adversa. Después de la probabilidad binomial se da la probabilidad de Poisson empleando $\lambda = Np = (2\,000)(0.001) = 2$. La aproximación de Poisson al parecer es en extremo cercana a la probabilidad binomial.

```
MTB > pdf 3;
SUBC> binomial 2000.001.
```

Función de probabilidad de densidad

```
Binomial con n = 2000 y p = 0.001
    x     P( X = x)
  3.0       0.1805
```

```
MTB > pdf 3;
SUBC> poisson 2.
```

Función de probabilidad de densidad

```
Poisson con mu = 2
     x    P( X = x)
  3.00       0.1804
```

b) La probabilidad de que más de dos individuos tengan una reacción adversa se obtiene de $1 - P(X \leq 2)$. El siguiente resultado de MINITAB da como probabilidad de que $X \leq 2$ el resultado 0.6767 usando tanto la distribución binomial como la distribución de Poisson. La probabilidad de que más de 2 tengan una reacción adversa es $1 - 0.6767 = 0.3233$.

```
MTB > cdf 2;
SUBC> binomial 2000.001.
```

Función de distribución acumulada

```
Binomial con n = 2000 y p = 0.001
    x     P( X <= x)
  2.0       0.6767
MTB > cdf 2;
SUBC> poisson 2.
```

Función de distribución acumulada

```
Poisson con mu = 2
    x     P( X <= x)
  2.0       0.6767
```

7.29 Una distribución de Poisson está dada por

$$p(X) = \frac{(0.72)^X e^{-0.72}}{X!}$$

Hallar: *a*) $p(0)$, *b*) $p(1)$, *c*) $p(2)$ y *d*) $p(3)$.

SOLUCIÓN

a) $$p(0) = \frac{(0.72)^0 e^{-0.72}}{0!} = \frac{(1)\,e^{-0.72}}{1} = e^{-0.72} = 0.4868$$ usando el apéndice VIII

b) $$p(1) = \frac{(0.72)^1 e^{-0.72}}{1!} = (0.72)e^{-0.72} = (0.72)(0.4868) = 0.3505$$

c) $$p(2) = \frac{(0.72)^2 e^{-0.72}}{2!} = \frac{(0.5184)e^{-0.72}}{2} = (0.2592)(0.4868) = 0.1262$$

Otro método

$$p(2) = \frac{0.72}{2}\,p(1) = (0.36)(0.3505) = 0.1262$$

d) $$p(3) = \frac{(0.72)^3 e^{-0.72}}{3!} = \frac{0.72}{3}\,p(2) = (0.24)(0.1262) = 0.0303$$

LA DISTRIBUCIÓN MULTINOMIAL

7.30 Una caja contiene 5 pelotas rojas, 4 pelotas blancas y 3 pelotas azules. De la caja se extrae al azar una pelota, se anota su color y se devuelve a la caja. Hallar la probabilidad de que de 6 pelotas extraídas de esta manera, 3 sean rojas, 2 sean blancas y 1 sea azul.

SOLUCIÓN

Pr{roja en cualquier extracción} $= \frac{5}{12}$, Pr{blanca en cualquier extracción} $= \frac{4}{12}$, y Pr{azul en cualquier extracción} $= \frac{3}{12}$; por lo tanto,

$$\Pr\{\text{3 sean rojas, 2 sean blancas, 1 sea azul}\} = \frac{6!}{3!2!1!}\left(\frac{5}{12}\right)^3\left(\frac{4}{12}\right)^2\left(\frac{3}{12}\right)^1 = \frac{625}{5\,184}$$

AJUSTE DE DATOS MEDIANTE DISTRIBUCIONES TEÓRICAS

7.31 Ajustar una distribución binomial a los datos del problema 2.17.

SOLUCIÓN

Se tiene Pr{X caras en un lanzamiento de cinco monedas} $= p(X) = \binom{5}{X} p^X q^{5-X}$, donde p y q son las probabilidades respectivas de cara y de cruz en un lanzamiento de una moneda. De acuerdo con el problema 7.11*a*), el número medio de caras es $\mu = Np = 5p$. En la distribución de frecuencias reales (u observadas), la cantidad media de caras es

$$\frac{\sum fX}{\sum f} = \frac{(38)(0) + (144)(1) + (342)(2) + (287)(3) + (164)(4) + (25)(5)}{1\,000} = \frac{2\,470}{1\,000} = 2.47$$

Igualando las medias teórica y real, $5p = 2.47$, o bien $p = 0.494$. Por lo tanto, la distribución binomial ajustada está dada por $p(X) = \binom{5}{X}(0.494)^X(0.506)^{5-X}$.

En la tabla 7.4 se enumeran estas probabilidades así como las frecuencias esperadas (teóricas) y las frecuencias reales. El ajuste parece ser bueno.

Tabla 7.4

Número de caras (X)	Pr{X caras}	Frecuencias esperadas	Frecuencias observadas
0	0.0332	33.2 o bien 33	38
1	0.1619	161.9 o bien 162	144
2	0.3162	316.2 o bien 316	342
3	0.3087	308.7 o bien 309	287
4	0.1507	150.7 o bien 151	164
5	0.0294	29.4 o bien 29	25

7.32 Usar la prueba de Kolmogorov-Smirnov de MINITAB para probar la normalidad de los datos de la tabla 7.5. Los datos representan el tiempo en horas por semana que 30 estudiantes universitarios usan su teléfono celular.

Tabla 7.5

16	17	15
14	14	16
12	16	12
13	11	14
10	15	14
13	15	16
17	13	15
14	18	14
11	12	13
13	15	12

SOLUCIÓN

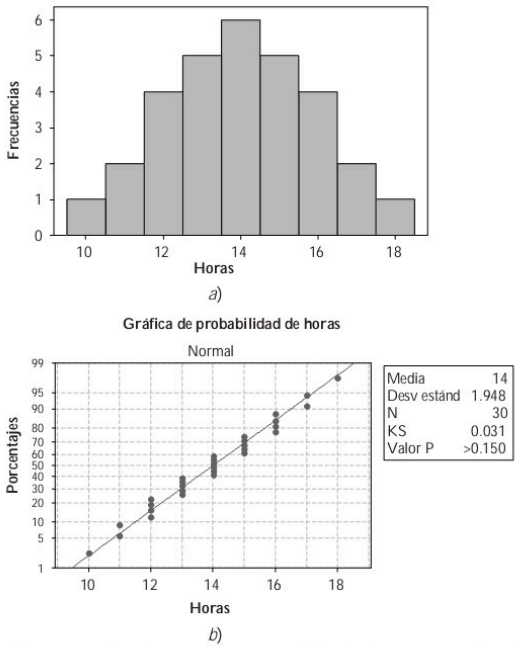


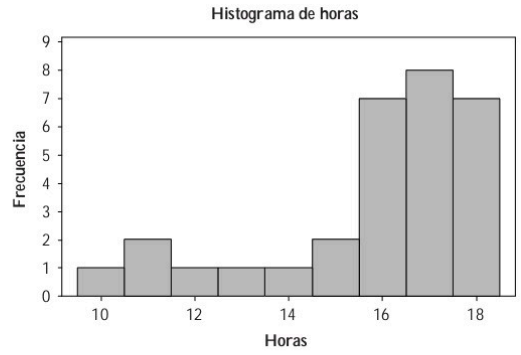
Figura 7-15 Prueba de Kolmogorov-Smirnov para normalidad: datos normales. *a*) Histograma que muestra un conjunto de datos distribuidos normalmente; *b*) la prueba de Kolmogorov-Smirnov para normalidad indica un valor $p > 0.150$ para normalidad.

El histograma de la figura 7-15*a*) indica que los datos de esta encuesta están distribuidos normalmente. La prueba de Kolgomorov-Smirnov también indica que los datos muestrales provienen de una población distribuida normalmente. La mayoría de los especialistas en estadística recomiendan que si el valor p es menor que 0.05, entonces se rechace la hipótesis de normalidad. Aquí el valor p es > 0.15.

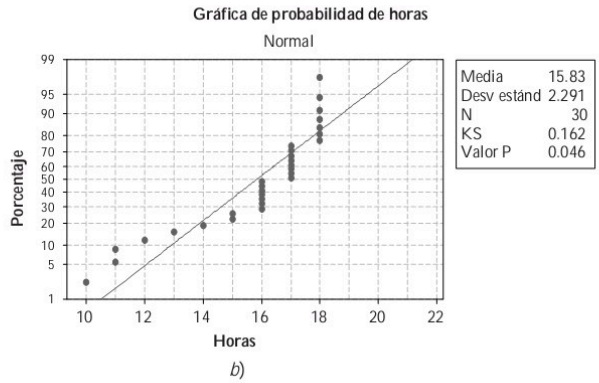
7.33 Usar la prueba de Kolgomorov-Smirnov de MINITAB para probar la normalidad de los datos presentados en la tabla 7.6. Estos datos son el tiempo en horas, por semana, que emplean su teléfono celular 30 estudiantes universitarios.

Tabla 7.6

18	16	11
17	12	13
17	17	17
16	18	17
16	17	17
16	18	15
18	18	16
16	16	17
14	18	15
11	18	10



a)



b)

Figura 7-16 Prueba de Kolgomorov-Smirnov para normalidad; valor $p = 0.046$, datos no normales. *a*) El histograma muestra un conjunto de datos sesgado a la izquierda; *b*) la prueba de Kolgomorov-Smirnov para normalidad indica carencia de normalidad.

La mayoría de los especialistas en estadística recomienda que si el valor p es menor que 0.05 se rechace la hipótesis de normalidad. En este caso el valor p es menor a 0.05.

7.34 En la tabla 7.7 se presenta el número de días, f, en el que ocurrieron X accidentes automovilísticos en una ciudad durante un periodo de 50 días. Ajustar a estos datos una distribución de Poisson.

Tabla 7.7

Cantidad de accidentes (X)	Cantidad de días (f)
0	21
1	18
2	7
3	3
4	1
	Total 50

SOLUCIÓN

La cantidad media de accidentes es

$$\lambda = \frac{\sum fX}{\sum f} = \frac{(21)(0) + (18)(1) + (7)(2) + (3)(3) + (1)(4)}{50} = \frac{45}{50} = 0.90$$

Por lo tanto, de acuerdo con la distribución de Poisson,

$$\Pr\{X \text{ accidentes}\} = \frac{(0.90)^X e^{-0.90}}{X!}$$

En la tabla 7.8 se dan las probabilidades, de 0, 1, 2, 3 y 4 accidentes, obtenidas con esta distribución de probabilidad de Poisson, así como las cantidades teóricas o esperadas de días con X accidentes (obtenidas multiplicando las probabilidades respectivas por 50). Para facilitar la comparación, en la columna 4 se dan nuevamente las cantidades de días de la tabla 7.7.

Obsérvese que el ajuste de la distribución de Poisson a los datos dados es bueno.

Tabla 7.8

Cantidad de accidentes (X)	$\Pr\{X \text{ accidentes}\}$	Cantidad esperada de días	Cantidad real de días
0	0.4066	23.33 o bien 20	21
1	0.3659	18.30 o bien 18	18
2	0.1647	8.24 o bien 8	7
3	0.0494	2.47 o bien 2	3
4	0.0111	0.56 o bien 1	1

En una distribución de Poisson, la varianza es $\sigma^2 = \lambda$. Calculando la varianza a partir de la distribución dada se obtiene 0.97, lo cual al compararlo con el valor 0.90 de λ resulta favorable, y esto puede tomarse como una evidencia más de que la distribución de Poisson es adecuada para aproximar los datos muestrales.

PROBLEMAS SUPLEMENTARIOS

LA DISTRIBUCIÓN BINOMIAL

7.35 Evaluar *a*) 7!, *b*) 10!/(6!4!), *c*), $\binom{9}{5}$ *d*) $\binom{11}{8}$ y *e*) $\binom{6}{1}$.

7.36 Expandir: *a*) $(q+p)^7$ y *b*) $(q+p)^{10}$.

7.37 Hallar la probabilidad de que al lanzar una moneda seis veces se obtengan: *a*) 0, *b*) 1, *c*) 2, *d*) 3, *e*) 4, *f*) 5 y *g*) 6 caras. *h*) Emplear MINITAB para construir una distribución de probabilidad para $X =$ número de caras en seis lanzamientos de una moneda.

7.38 Hallar la probabilidad de que en un solo lanzamiento de seis monedas se obtengan: *a*) 2 o más caras y *b*) menos de 4 caras. *c*) Usar EXCEL para hallar las respuestas de los incisos *a*) y *b*).

7.39 Si X denota el número de caras en un solo lanzamiento de cuatro monedas, encontrar: *a*) $\Pr\{X = 3\}$, *b*) $\Pr\{X < 2\}$, *c*) $\Pr\{X \leq 2\}$ y *d*)$\Pr\{1 < X \leq 3\}$.

7.40 De 800 familias con 5 hijos cada una, ¿cuántas se esperaría que tuvieran: *a*) 3 niños, *b*) 5 niñas y *c*) 2 o 3 niños? Supónganse iguales probabilidades para niños y niñas.

7.41 Encontrar la probabilidad de obtener, en dos lanzamientos de un par de dados, la suma 11: *a*) una vez y *b*) dos veces.

7.42 ¿Cuál es la probabilidad de obtener 9 una sola vez en tres lanzamientos de un par de dados?

7.43 Hallar la probabilidad de adivinar, correctamente, por lo menos 6 de 10 respuestas en un examen de verdadero y falso.

7.44 Un vendedor de seguros vende pólizas a 5 hombres, todos de la misma edad y con buena salud. De acuerdo con las tablas actuariales, la probabilidad de que un hombre de esta edad esté vivo en 30 años es $\frac{2}{3}$. Encontrar la probabilidad de que en 30 años estén vivos: *a*) los 5 hombres, *b*) por lo menos 3 de estos hombres, *c*) sólo 2 de estos hombres y *d*) por lo menos 1 de ellos. *e*) Usar EXCEL para responder los incisos del *a*) al *d*).

7.45 Calcular: *a*) la media, *b*) la desviación estándar, *c*) el coeficiente momento de sesgo y *d*) el coeficiente momento de curtosis de la distribución binomial en la que $p = 0.7$ y $N = 60$. Interpretar los resultados.

7.46 Mostrar que si una distribución binomial en la que $N = 100$ es simétrica, su coeficiente momento de curtosis es 2.98.

7.47 Evaluar: *a*) $\sum (X-\mu)^3 p(X)$ y *b*) $\sum (X-\mu)^4 p(X)$ para una distribución binomial.

7.48 Probar las fórmulas (*1*) y (*2*) al principio de este capítulo respecto a los coeficientes momento de sesgo y de curtosis.

LA DISTRIBUCIÓN NORMAL

7.49 En un examen de estadística, la puntuación media es 78 y la desviación estándar es 10.

a) Determinar las puntuaciones estándar de dos estudiantes cuyas calificaciones fueron 93 y 62, respectivamente.

b) Determinar las calificaciones de dos estudiantes cuyas puntuaciones estándar fueron -0.6 y 1.2, respectivamente.

7.50 Encontrar: *a*) la media y *b*) la desviación estándar de las calificaciones obtenidas en un examen en el que 70 y 88 corresponden a las puntuaciones estándar -0.6 y 1.4, respectivamente.

7.51 Hallar el área bajo la curva normal entre: *a*) $z = -1.20$ y $z = 2.40$; *b*) $z = 1.23$ y $z = 1.87$, y *c*) $z = -2.35$ y $z = -0.50$. *d*) Resolver los incisos del *a*) al *c*) empleando EXCEL.

7.52 Hallar el área bajo la curva normal: *a*) a la izquierda de $z = -1.78$, *b*) a la izquierda de $z = 0.56$, *c*) a la derecha de $z = -1.45$, *d*) correspondiente a $z \geq 2.16$, *e*) correspondiente a $-0.80 \leq z \leq 1.53$ y *f*) a la izquierda de $z = -2.52$ y a la derecha de $z = 1.83$. *g*) Resolver los incisos del *a*) al *f*) usando EXCEL.

7.53 Si z está distribuida normalmente con media 0 y varianza 1, hallar: *a*) $\Pr\{z \geq -1.64\}$, *b*) $\Pr\{-1.96 \leq z \leq 1.96\}$ y *c*) $\Pr\{|z| \geq 1\}$.

7.54 Hallar el valor de z tal que: *a*) el área a la derecha de z sea 0.2266, *b*) el área a la izquierda de z sea 0.0314, *c*) el área entre -0.23 y z sea 0.5722, *d*) el área entre 1.15 y z sea 0.0730 y *e*) el área entre $-z$ y z sea 0.9000.

7.55 Encontrar z_1 si $\Pr\{z \geq z_1\} = 0.84$, donde z está distribuida normalmente con media 0 y varianza 1.

7.56 Empleando el apéndice I, encontrar las ordenadas en la curva normal correspondientes a: *a*) $z = 2.25$, *b*) $z = -0.32$ y *c*) $z = -1.18$. *d*) Resolver los incisos del *a*) al *c*) empleando EXCEL.

7.57 Las estaturas de hombres adultos tienen una distribución normal cuya media es 70 in y cuya desviación estándar es 3 in. *a*) ¿Qué porcentaje mide menos de 65 in? *b*) ¿Qué porcentaje mide más de 72 in? *c*) ¿Qué porcentaje está entre 68 y 73 in?

7.58 Las cantidades gastadas, por determinado grupo de edad, en la compra de artículos en línea tienen una distribución normal cuya media es \$125 y cuya desviación estándar es \$25. *a*) ¿Qué porcentaje gasta más de \$175? *b*) ¿Qué porcentaje gasta entre \$100 y \$150? *c*) ¿Qué porcentaje gasta menos de \$50?

7.59 En un examen final la calificación media es 72 y la desviación estándar es 9. Los estudiantes que forman parte del 10% superior obtienen A como nota. ¿Cuál es la calificación mínima para obtener A?

7.60 Si un conjunto de medidas tiene una distribución normal, ¿qué porcentaje de las medidas difiere de la media en: *a*) más de media desviación estándar y *b*) menos de tres cuartos de desviación estándar?

7.61 Si $\bar{X}$ es la media y s es la desviación estándar de un conjunto de mediciones distribuidas normalmente, ¿qué porcentaje de las mediciones: *a*) está dentro del rango $\bar{X} \pm 2s$, *b*) están fuera del rango $\bar{X} \pm 1.2s$ y *c*) son mayores que $\bar{X} - 1.5s$?

7.62 En el problema 7.61 encontrar la constante a tal que el porcentaje de casos: *a*) dentro del rango $\bar{X} \pm as$ sea 75% y *b*) menores que $\bar{X} - as$ sea 22%.

APROXIMACIÓN NORMAL A LA DISTRIBUCIÓN BINOMIAL

7.63 Encontrar la probabilidad de que en 200 lanzamientos de una moneda se obtengan: *a*) entre 80 y 120 caras inclusive, *b*) menos de 90 caras, *c*) menos de 85 o más de 115 caras y *d*) exactamente 100 caras.

7.64 Encontrar la probabilidad de que en un examen de verdadero o falso un estudiante adivine correctamente las respuestas de: *a*) 12 de 20 preguntas o más y *b*) 24 de 40 preguntas o más.

7.65 De los tornillos que se producen con una máquina, 10% está defectuoso. Encontrar la probabilidad de que en una muestra aleatoria de 400 tornillos producidos con esta máquina: *a*) cuando mucho 30, *b*) entre 30 y 50, *c*) entre 35 y 45 y *d*) 55 o más de los tornillos estén defectuosos.

7.66 Encontrar la probabilidad de obtener más de 25 sietes en 100 lanzamientos de un par de dados.

DISTRIBUCIÓN DE POISSON

7.67 Si 3% de los bulbos eléctricos fabricados por una empresa están defectuosos, encontrar la probabilidad de que en una muestra de 100 bulbos: *a*) 0, *b*) 1, *c*) 2, *d*) 3, *e*) 4 y *f*) 5 bulbos estén defectuosos.

7.68 En el problema 7.67, hallar la probabilidad de que: *a*) más de 5, *b*) entre 1 y 3 y *c*) 2 o menos bulbos estén defectuosos.

7.69 Una bolsa contiene 1 canica roja y 7 canicas blancas. De la bolsa se extrae una canica y se observa su color. Después se regresa la canica a la bolsa y el contenido de la bolsa se mezcla bien. Usando: *a*) la distribución binomial y *b*) la aproximación de Poisson a la distribución binomial, encontrar la probabilidad de que en 8 extracciones se extraiga exactamente 3 veces una canica roja.

7.70 Según la oficina de estadística del Departamento de Salud de Estados Unidos, en ese país la cantidad anual de ahogados accidentalmente es 3 por 100 000 habitantes. Encontrar la probabilidad de que en una ciudad en que la población es de 200 000 habitantes haya anualmente: *a*) 0, *b*) 2, *c*) 6, *d*) 8, *e*) entre 4 y 8, y *f*) menos de 3 ahogados en forma accidental.

7.71 En una empresa, la cantidad promedio de llamadas que llegan al conmutador entre las 2 y las 4 de la tarde es 2.5 por minuto. Encontrar la probabilidad de que en determinado minuto haya: *a*) 0, *b*) 1, *c*) 2, *d*) 3, *e*) 4 o menos, y *f*) más de 6 llamadas.

LA DISTRIBUCIÓN MULTINOMIAL

7.72 Un dado se lanza seis veces. Hallar la probabilidad de que se obtengan: *a*) un 1, dos 2 y tres 3, y *b*) una vez cada lado.

7.73 Una caja contiene una gran cantidad de canicas rojas, blancas, azules y amarillas en la proporción 4 : 3 : 2 : 1, respectivamente. Encontrar la probabilidad de que en diez extracciones los colores de las canicas sean: *a*) 4 rojas, 3 blancas, 2 azules y una amarilla y *b*) 8 rojas y 2 amarillas.

7.74 Encontrar la probabilidad de que en cuatro lanzamientos de un dado no se obtengan 1, 2 ni 3.

AJUSTE DE DATOS A DISTRIBUCIONES TEÓRICAS

7.75 Ajustar la distribución binomial a los datos de la tabla 7.9.

Tabla 7.9

X	0	1	2	3	4
f	30	62	46	10	2

7.76 En una encuesta realizada a estudiantes de educación media se determinaron las horas de ejercicio que practican por semana. Usando STATISTIX, construir un histograma con los datos. Empleando la prueba de Shapiro-Wilk de STATISTIX, determinar si los datos provienen de una distribución normal. Los datos aparecen en la tabla 7.10.

Tabla 7.10

5	10	2	3	2
5	5	1	3	15
1	2	20	3	1
4	4	4	3	5

7.77 Con los datos de la tabla 7.5 del problema 7.32, construir un histograma empleando STATISTIX. Empleando la prueba de Shapiro-Wilk de STATISTIX, determinar si los datos provienen de una distribución normal.

7.78 Las puntuaciones de examen de la tabla 7.11 siguen una distribución en forma de U, que es exactamente lo opuesto a una distribución normal. Con estos datos, construir un histograma empleando STATISTIX. Usando la prueba de Shapiro-Wilk de STATISTIX, determinar si los datos provienen de una distribución normal.

Tabla 7.11

20	90	10
40	90	20
80	70	50
70	40	90
90	70	10
60	30	80
10	20	30
30	10	20
10	80	90
60	50	80

7.79 Emplee los datos de la tabla 7.11. Usando la prueba de Anderson-Darling y la de Ryan-Joiner de MINITAB, determinar si los datos provienen de una distribución normal.

7.80 Los datos de la tabla 7.12 provienen de 10 cuerpos de la armada prusiana y corresponden a un periodo de 20 años (1875 a 1894). Estos datos muestran la cantidad de muertes, por año, debidas a patadas de caballo. Ajustar una distribución de Poisson a los datos.

Tabla 7.12

X	0	1	2	3	4
f	109	65	22	3	1

TEORÍA ELEMENTAL DEL MUESTREO

8

TEORÍA DEL MUESTREO

La *teoría del muestreo* es el estudio de la relación que existe entre una población y las muestras que se obtienen de esa población. La teoría del muestreo se emplea en muchos contextos. Por ejemplo, en la *estimación* de cantidades poblacionales desconocidas (como la media y la varianza poblacionales), a las que se les conoce como *parámetros poblacionales* o simplemente *parámetros*, a partir de las correspondientes cantidades muestrales (como la media y la varianza muestrales), a menudo conocidas como *estadísticos muestrales* o simplemente *estadísticos*. El problema de la estimación se estudia en el capítulo 9.

La teoría del muestreo también sirve para determinar si las diferencias que se observan entre dos muestras se deben a variaciones casuales o si son diferencias realmente significativas. Tales preguntas surgen, por ejemplo, al probar un nuevo suero para el tratamiento de una enfermedad o cuando se tiene que decidir si un proceso de producción es mejor que otro. Para responder a estas preguntas se usan las llamadas *pruebas de significancia o de hipótesis*, fundamentales en la *teoría de decisiones*. Estos temas se tratan en el capítulo 10.

En general, al estudio de las inferencias que se hacen acerca de una población, empleando muestras obtenidas de ella, y de las indicaciones de la exactitud de tales inferencias, mediante el uso de la teoría de la probabilidad, es a lo que se le llama *inferencia estadística*.

MUESTRAS ALEATORIAS Y NÚMEROS ALEATORIOS

Para que las conclusiones que se obtienen empleando la teoría del muestreo y la inferencia estadística sean válidas, las muestras deben elegirse de manera que sean *representativas* de la población. Al estudio de los métodos de muestreo y de los problemas relacionados con ellos se le conoce como *diseño de experimentos*.

Una manera de obtener una muestra representativa es mediante un proceso llamado *muestreo aleatorio*, mediante el cual cada uno de los miembros de la población tiene la misma posibilidad de ser incluido en la muestra. Una técnica para obtener una muestra aleatoria consiste en asignarle, a cada miembro de la población, un número, escribir estos números en pedazos pequeños de papel, colocarlos en una urna y después extraer los números de la urna, teniendo cuidado de mezclar muy bien antes de cada extracción. Una alternativa a este método es usar una tabla de *números aleatorios* (ver el apéndice IX), la cual se construye especialmente para este fin. Ver el problema 8.6.

MUESTREO CON REPOSICIÓN Y SIN ELLA

Si se extrae un número de una urna, antes de extraer otro, el número puede ser devuelto a la urna (ser repuesto) o no. En el primer caso, el número puede ser extraído varias veces, en tanto que en el segundo caso sólo puede ser extraído una vez. A un muestreo en el que cada miembro de la población puede ser elegido más de una vez se le llama *muestreo con reposición*; en cambio, si sólo puede ser elegido una vez se llama *muestreo sin reposición*.

Una población puede ser finita o infinita. Por ejemplo, si de una urna que contiene 100 canicas se extraen sucesivamente 10 canicas sin reposición, se está muestreando una población finita; en cambio, si se lanza una moneda 50 veces y se cuenta la cantidad de caras, se está muestreando de una población infinita.

Una población finita que se muestrea con reposición puede considerarse teóricamente infinita, ya que se puede extraer cualquier cantidad de muestras sin agotar la población. Para fines prácticos, cuando se muestrea de una población finita pero muy grande, se puede considerar que el muestreo se hace de una población infinita.

DISTRIBUCIONES MUESTRALES

Considérense todas las muestras de tamaño N que pueden extraerse de determinada población (ya sea con reposición o sin ella). Para cada muestra se pueden calcular diversos estadísticos (como media o desviación estándar), los cuales variarán de una muestra a otra. De esta manera se obtiene una distribución del estadístico de que se trate, a la que se le llama *distribución muestral*.

Por ejemplo, si el estadístico de que se trata es la media muestral, a la distribución que se obtiene se le llama *distribución muestral de las medias* o *distribución muestral de la media*. De igual manera se pueden obtener distribuciones muestrales de las desviaciones estándar, de las varianzas, de las medianas, de las proporciones, etcétera.

A cada distribución muestral se le puede calcular su media, su desviación estándar, etc. Así, se puede hablar de la media, de la desviación estándar, de la distribución muestral de las medias, etcétera.

DISTRIBUCIONES MUESTRALES DE MEDIAS

Supóngase que de una población finita de tamaño $N_p > N$ se extraen, sin reposición, todas las muestras posibles de tamaño N. Si se denota con $\mu_{\bar{X}}$ y $\sigma_{\bar{X}}$ respectivamente, a la media y a la desviación estándar de una distribución muestral de las medias, y con μ y σ, respectivamente, a la media y la desviación estándar poblacionales, entonces

$$\mu_{\bar{X}} = \mu \qquad \text{y} \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}\sqrt{\frac{N_p - N}{N_p - 1}} \tag{1}$$

Si la población es infinita, o si el muestreo se hace con reposición, las fórmulas anteriores se reducen a

$$\mu_{\bar{X}} = \mu \qquad \text{y} \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} \tag{2}$$

Si el valor de N es grande ($N \geq 30$), la distribución muestral de las medias es aproximadamente normal con media $\mu_{\bar{X}}$ y desviación estándar $\sigma_{\bar{X}}$, independientemente de la población (siempre y cuando la media y la varianza poblacionales sean finitas y el tamaño de la población sea por lo menos el doble del tamaño de la muestra). Si la población es infinita, este resultado es un caso especial del *teorema del límite central* de la teoría avanzada de la probabilidad, el cual muestra que la exactitud de la aproximación aumenta a medida que N aumenta. Esto suele indicarse diciendo que la distribución muestral es *asintóticamente normal*.

Si la población está distribuida normalmente, la distribución muestral de las medias también es normal aun cuando el valor de N sea pequeño (es decir, $N < 30$).

DISTRIBUCIONES MUESTRALES DE PROPORCIONES

Supóngase que una población sea infinita y que la probabilidad de ocurrencia de un evento (llamada éxito) es p, y que la probabilidad de no ocurrencia del evento es $q = 1 - p$. La población puede ser, por ejemplo, la de los lanzamientos de una moneda, en los que la probabilidad del evento "cara" es $p = \frac{1}{2}$. Considérense todas las posibles muestras de tamaño N extraídas de esta población, y para cada muestra determínese la proporción P de éxitos. En el caso de una moneda, P es la proporción de caras en N lanzamientos. De esta manera se obtiene una *distribución muestral de las proporciones* cuya media μ_P y cuya desviación estándar σ_P están dadas por

$$\mu_P = p \qquad \text{y} \qquad \sigma_P = \sqrt{\frac{pq}{N}} = \sqrt{\frac{p(1-p)}{N}} \tag{3}$$

que se pueden obtener de la ecuación (*2*) sustituyendo $\mu = p$ y $\sigma = \sqrt{pq}$. Si el valor de N es grande ($N \geq 30$), esta distribución muestral es aproximadamente normal. Obsérvese que la población está *distribuida en forma binomial*.

Las ecuaciones (*3*) también son válidas para poblaciones finitas si el muestreo se hace con reposición. En el caso de poblaciones finitas en las que el muestreo se hace sin reposición, las ecuaciones (*3*) se sustituyen por las ecuaciones (*1*) con $\mu = p$ y $\sigma = \sqrt{pq}$.

Obsérvese que las ecuaciones (*3*) pueden obtenerse más fácilmente dividiendo entre N, la media y la desviación estándar (Np y $\sqrt{Npq}$) de la distribución binomial (ver capítulo 7).

DISTRIBUCIONES MUESTRALES DE DIFERENCIAS Y SUMAS

Se supone que se tienen dos poblaciones. Para cada muestra de tamaño N_1 tomada de la primera población se calcula un estadístico S_1, con lo que se obtiene una distribución muestral de este estadístico S_1, cuya media y desviación estándar se denotan μ_{S1} y σ_{S1}, respectivamente. De igual manera, para cada muestra de tamaño N_2 tomada de la segunda población se calcula un estadístico S_2, con lo que se obtiene una distribución muestral de este estadístico S_2, cuya media y desviación estándar se denotan μ_{S2} y σ_{S2}, respectivamente. Con todas las posibles combinaciones de estas muestras de las dos poblaciones se obtiene una distribución de las diferencias, $S_1 - S_2$, a la que se le llama *distribución muestral de las diferencias de los estadísticos*. La media y la desviación estándar de esta distribución muestral se denotan, respectivamente, μ_{S1-S2} y σ_{S1-S2}, y están dadas por

$$\mu_{S1-S2} = \mu_{S1} - \mu_{S2} \qquad \text{y} \qquad \sigma_{S1-S2} = \sqrt{\sigma_{S1}^2 + \sigma_{S2}^2} \tag{4}$$

siempre y cuando las muestras elegidas no dependan, de manera alguna, una de la otra (es decir, las muestras sean *independientes*).

Si S_1 y S_2 son las medias muestrales de las dos poblaciones, a las que se les denota $\bar{X}_1$ y $\bar{X}_2$, respectivamente, entonces la distribución muestral de las diferencias de las medias está dada para poblaciones infinitas con media y desviación estándar (μ_1 y σ_1) y (μ_2 y σ_2), respectivamente, por

$$\mu_{\bar{X}1-\bar{X}2} = \mu_{\bar{X}1} - \mu_{\bar{X}2} = \mu_1 - \mu_2 \qquad \text{y} \qquad \sigma_{\bar{X}1-X2} = \sqrt{\sigma_{\bar{X}1}^2 + \sigma_{\bar{X}2}^2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} \tag{5}$$

usando las ecuaciones (*2*). Estas ecuaciones también son válidas para poblaciones finitas si el muestreo se hace con reposición. Para poblaciones finitas en las que el muestreo se haga sin reposición, se obtienen ecuaciones similares empleando las ecuaciones (*1*).

Tabla 8.1 Error estándar de distribuciones muestrales

Distribución muestral	Error estándar	Observaciones
Media	$\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{N}}$	Esta fórmula es válida tanto para muestras grandes como para muestras pequeñas. La distribución muestral de las medias se aproxima a una distribución normal cuando $N \geq 30$, aun cuando la población no sea normal. $\mu_{\bar{X}} = \mu$, la media poblacional, en todos los casos.
Proporciones	$\sigma_P = \sqrt{\dfrac{p(1-p)}{N}} = \sqrt{\dfrac{pq}{N}}$	La observación hecha para las medias también es válida en este caso. $\mu_P = p$, en todos los casos.
Desviaciones estándar	(*1*) $\sigma_s = \dfrac{\sigma}{\sqrt{2N}}$ (*2*) $\sigma_s = \sqrt{\dfrac{\mu_4 - \mu_2^2}{4N\mu_2}}$	Cuando $N \geq 100$, la distribución muestral de s es casi normal. σ_s está dada por (*1*) sólo si la población es normal (o aproximadamente normal). Si la población no es normal, se puede usar (*2*). Obsérvese que (*2*) se reduce a (*1*) cuando $\mu_2 = \sigma^2$ y $\mu_4 = 3\sigma^4$, lo que ocurre en poblaciones normales. Cuando $N \geq 100$, $\mu_s = \sigma$ muy aproximadamente.
Medianas	$\sigma_{\text{med}} = \sigma\sqrt{\dfrac{\pi}{2N}} = \dfrac{1.2533\sigma}{\sqrt{N}}$	Cuando $N \geq 30$, la distribución muestral de la mediana es casi normal. La fórmula dada es válida sólo si la población es normal (o aproximadamente normal). $\mu_{\text{med}} = \mu$
Primero y tercer cuartiles	$\sigma_{Q1} = \sigma_{Q3} = \dfrac{1.3626\sigma}{\sqrt{N}}$	La observación hecha para las medianas también es válida aquí. μ_{Q1} y μ_{Q3} son casi iguales al primero y tercer cuartiles de la población. Obsérvese que $\sigma_{Q2} = \sigma_{\text{med}}$
Deciles	$\sigma_{D1} = \sigma_{D9} = \dfrac{1.7094\sigma}{\sqrt{N}}$ $\sigma_{D2} = \sigma_{D8} = \dfrac{1.4288\sigma}{\sqrt{N}}$ $\sigma_{D3} = \sigma_{D7} = \dfrac{1.3180\sigma}{\sqrt{N}}$ $\sigma_{D4} = \sigma_{D6} = \dfrac{1.2680\sigma}{\sqrt{N}}$	Las observaciones hechas para las medianas también son válidas aquí. $\mu_{D1}, \mu_{D2}, \ldots$ son casi iguales al primero, segundo, ... deciles de la población. Obsérvese que $\sigma_{D5} = \sigma_{\text{med}}$.
Rangos semiintercuartílicos	$\sigma_Q = \dfrac{0.7867\sigma}{\sqrt{N}}$	Las observaciones hechas para las medianas también son válidas aquí. μ_Q es casi igual al rango semiintercuartil poblacional.
Varianzas	(*1*) $\sigma_{S^2} = \sigma^2\sqrt{\dfrac{2}{N}}$ (*2*) $\sigma_{S^2} = \sqrt{\dfrac{\mu_4 - \dfrac{N-3}{N-1}\mu_2^2}{N}}$	Las observaciones hechas para la desviación estándar también son válidas aquí. Obsérvese que si la población es normal (*2*) da (*1*). $\mu_{S^2} = \sigma^2(N-1)/N$, que es casi igual a σ^2 cuando N es grande.
Coeficiente de variación	$\sigma_V = \dfrac{v}{\sqrt{2N}}\sqrt{1+2v^2}$	Aquí $v = \sigma/\mu$ es el coeficiente de variación poblacional. La fórmula dada es válida para poblaciones normales (o casi normales) y $N \geq 100$.

Pueden obtenerse resultados semejantes para las distribuciones muestrales de las diferencias entre las proporciones de dos poblaciones distribuidas en forma binomial con parámetros (p_1, q_1) y (p_2, q_2), respectivamente. En este caso, S_1 y S_2 son proporción de éxitos, P_1 y P_2, y las ecuaciones (*4*) se transforman en

$$\mu_{P1-P2} = \mu_{P1} - \mu_{P2} = p_1 - p_2 \qquad \text{y} \qquad \sigma_{P1-P2} = \sqrt{\sigma_{P1}^2 + \sigma_{P2}^2} = \sqrt{\frac{p_1 q_1}{N_1} + \frac{p_2 q_2}{N_2}} \tag{6}$$

Si N_1 y N_2 son grandes ($N_1, N_2 \geq 30$), la distribución muestral de diferencias entre medias o proporciones se aproximan mucho a una distribución normal.

Algunas veces se necesita la *distribución muestral de la suma de estadísticos.* La media y la desviación estándar de estas distribuciones están dadas por

$$\mu_{S1+S2} = \mu_{S1} + \mu_{S2} \qquad \text{y} \qquad \sigma_{S1+S2} = \sqrt{\sigma_{S1}^2 + \sigma_{S2}^2} \tag{7}$$

suponiendo que las muestras sean *independientes.*

ERRORES ESTÁNDAR

A la desviación estándar de la distribución muestral de un estadístico suele conocérsele como su *error estándar.* En la tabla 8.1 se enumeran los errores estándar de distribuciones muestrales de varios estadísticos, suponiendo que el muestreo es un muestreo aleatorio de una población infinita (o muy grande) o de una población finita pero hecho el muestreo con reposición. Se presentan también algunas observaciones especiales en las que se dan las condiciones bajo las cuales son válidas las fórmulas, así como otras observaciones pertinentes.

Las cantidades μ, σ, p, μ_r y $\bar{X}$, s, P, m_r denotan media, desviación estándar, proporción y el r-ésimo momento respecto a la media, poblacionales y muestrales, respectivamente.

Se hace notar que si el tamaño N de la muestra es suficientemente grande, la distribución muestral es normal o casi normal. A esto se debe que estos métodos se conozcan como *métodos para muestras grandes.* Cuando $N < 30$, a las muestras se les llama *pequeñas.* La teoría de las muestras *pequeñas* o *teoría del muestreo exacto*, como se le llama algunas veces, se estudia en el capítulo 11.

Cuando los parámetros poblacionales, por ejemplo, σ, p o bien μ_r no se conocen, pueden estimarse con bastante exactitud a partir de sus estadísticos muestrales, s (o bien $\hat{s} = \sqrt{N/(N-1)}s$), P y m_r, siempre y cuando las muestras sean suficientemente grandes.

DEMOSTRACIONES DE LA TEORÍA ELEMENTAL DEL MUESTREO EMPLEANDO SOFTWARE

EJEMPLO 1

En una población grande se define la siguiente variable aleatoria. X es la cantidad de computadoras por hogar; X está distribuida de manera uniforme, es decir, $p(x) = 0.25$ para $x = 1, 2, 3$ y 4. En otras palabras, 25% de los hogares tiene 1 computadora; 25% tiene 2 computadoras; 25% tiene tres computadoras y 25% tiene 4 computadoras. La media de X es $\mu = \Sigma x p(x) = 0.25 + 0.5 + 0.75 + 1 = 2.5$. La varianza de X es $\sigma^2 = \Sigma x^2 p(x) - \mu^2 = 0.25 + 1 + 2.25 + 4 - 6.25 = 1.25$. Entonces, la cantidad media de computadoras por hogar es 2.5 y la varianza de la cantidad de computadoras por hogar es 1.25.

EJEMPLO 2

Para enumerar todas las muestras, tomadas con reposición, de dos hogares puede usarse MINITAB. La hoja de cálculo se verá como la que se presenta en la tabla 8.2. Las 16 muestras aparecen en C1 y C2, y la media de cada una de ellas en C3. Como la población está distribuida de manera uniforme, la probabilidad de cada media muestral es 1/16. Resumiendo, en las columnas C4 y C5 se da la distribución de probabilidad.

Obsérvese que $\mu_{\bar{x}} = \Sigma \bar{x} p(\bar{x}) = 1(0.0625) + 1.5(0.1250) + \cdots + 4(0.0625) = 2.5$. Como se ve $\mu_{\bar{x}} = \mu$. Además, $\sigma_{\bar{x}}^2 = \Sigma \bar{x}^2 p(\bar{x}) - \mu_{\bar{x}}^2 = 1(0.0625) + 2.25(0.1250) + \cdots + 16(0.0625) - 6.25 = 0.625$ con lo que $\sigma_{\bar{x}}^2 = (\sigma^2/2)$. Empleando MINITAB para dibujar la gráfica de la distribución de probabilidad de x barra se obtiene el resultado que se muestra en la figura 8-1. (Obsérvese que $\bar{X}$ y x barra se usan indistintamente.)

Tabla 8.2

C1 hogar 1	C2 hogar 2	C3 media	C4 x barra	C5 p(x barra)
1	1	1.0	1.0	0.0625
1	2	1.5	1.5	0.1250
1	3	2.0	2.0	0.1875
1	4	2.5	2.5	0.2500
2	1	1.5	3.0	0.1875
2	2	2.0	3.5	0.1250
2	3	2.5	4.0	0.0625
2	4	3.0		
3	1	2.0		
3	2	2.5		
3	3	3.0		
3	4	3.5		
4	1	2.5		
4	2	3.0		
4	3	3.5		
4	4	4.0		

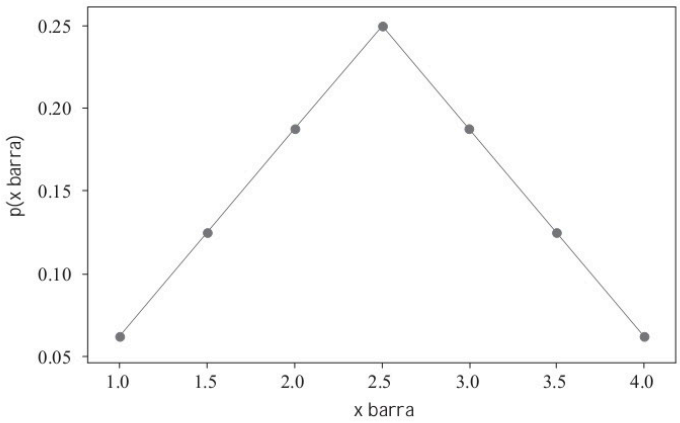


Figura 8-1 Gráfica de *p*(x barra) *vs.* x barra.

PROBLEMAS RESUELTOS

DISTRIBUCIÓN MUESTRAL DE LAS MEDIAS

8.1 Una población consta de los cinco números 2, 3, 6, 8 y 11. Considerar todas las muestras de tamaño 2 que pueden extraerse de esta población con reposición. Encontrar: *a*) la media de la población, *b*) la desviación estándar de la población, *c*) la media de la distribución muestral de las medias y *d*) la desviación estándar de la distribución muestral de las medias (es decir, el error estándar de las medias).

SOLUCIÓN

a)
$$\mu = \frac{2+3+6+8+11}{5} = \frac{30}{5} = 6.0$$

b)
$$\sigma^2 = \frac{(2-6)^2+(3-6)^2+(6-6)^2+(8-6)^2+(11-6)^2}{5} = \frac{16+9+0+4+25}{5} = 10.8$$

y $\sigma = 3.29$.

c) Existen $5(5) = 25$ muestras de tamaño 2 que pueden extraerse con reposición (ya que a cada uno de los cinco números de la primera extracción le corresponden cada uno de los cinco números de la segunda extracción). Así, se tiene

(2, 2)	(2, 3)	(2, 6)	(2, 8)	(2, 11)
(3, 2)	(3, 3)	(3, 6)	(3, 8)	(3, 11)
(6, 2)	(6, 3)	(6, 6)	(6, 8)	(6, 11)
(8, 2)	(8, 3)	(8, 6)	(8, 8)	(8, 11)
(11, 2)	(11, 3)	(11, 6)	(11, 8)	(11, 11)

Las medias muestrales correspondientes son

2.0	2.5	4.0	5.0	6.5
2.5	3.0	4.5	5.5	7.0
4.0	4.5	6.0	7.0	8.5
5.0	5.5	7.0	8.0	9.5
6.5	7.0	8.5	9.5	11.0

(8)

y la media de la distribución muestral de las medias es

$$\mu_{\bar{X}} = \frac{\text{suma de todas las medias muestrales de } (8)}{25} = \frac{150}{25} = 6.0$$

lo que ilustra que $\mu_{\bar{X}} = \mu$.

d) La varianza $\sigma_{\bar{X}}^2$ de la distribución muestral de las medias se obtiene restándole 6 a cada una de las medias en (8), elevando cada resultado al cuadrado, sumando los 25 resultados obtenidos y dividiendo esta suma entre el 25. El resultado final es $\sigma_{\bar{X}}^2 = 135/25 = 5.40$ y por lo tanto $\sigma_{\bar{X}} = \sqrt{5.40} = 2.32$. Esto ilustra que en una población finita en la que se muestrea con reposición (o en una población infinita), $\sigma_{\bar{X}}^2 = \sigma^2/N$, ya que el lado derecho es $10.8/2 = 5.40$, que coincide con el valor anterior.

8.2 Resolver el problema 8.1 considerando que el muestreo se hace sin reposición.

SOLUCIÓN

Como en los incisos *a*) y *b*) del problema 8.1, $\mu = 6$ y $\sigma = 3.29$.

c) Existen $\binom{5}{2} = 10$ muestras de tamaño 2 que pueden ser extraídas sin reposición (esto significa que se extrae un número y después otro diferente al primero) de la población: (2, 3), (2, 6), (2, 8), (2, 11), (3, 6), (3, 8), (3, 11), (6, 8), (6, 11) y (8, 11). La extracción (2, 3) se considera igual a la (3, 2).

Las medias muestrales correspondientes son 2.5, 4.0, 5.0, 6.5, 4.5, 5.5, 7.0, 7.0, 8.5 y 9.5, y la media de la distribución muestral de las medias es

$$\mu_{\bar{X}} = \frac{2.5+4.0+5.0+6.5+4.5+5.5+7.0+7.0+8.5+9.5}{10} = 6.0$$

lo que ilustra que $\mu_{\bar{X}} = \mu$.

d) La varianza de la distribución muestral de las medias es

$$\sigma_{\bar{X}}^2 = \frac{(2.5 - 6.0)^2 + (4.0 - 6.0)^2 + (5.0 - 6.0)^2 + \cdots + (9.5 - 6.0)^2}{10} = 4.05$$

y $\sigma_{\bar{X}} = 2.01$. Esto ilustra que

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{N}\left(\frac{N_p - N}{N_p - 1}\right)$$

ya que el lado derecho es igual a

$$\frac{10.8}{2}\left(\frac{5-2}{5-1}\right) = 4.05$$

que es lo que se obtuvo antes.

8.3 Supóngase que las estaturas de 3 000 estudiantes del sexo masculino de una universidad tienen una distribución normal con media 68.0 pulgadas (in) y desviación estándar 3.0 in. Si se obtienen 80 muestras, cada una de 25 estudiantes, ¿cuáles serán la media y la desviación estándar esperadas de la distribución muestral de las medias si el muestreo se hace: *a*) con reposición y *b*) sin reposición?

SOLUCIÓN

El número de muestras de tamaño 25 que teóricamente pueden obtenerse de un grupo de 3 000 estudiantes, con reposición y sin ésta son, respectivamente $(3\,000)^{25}$ y $\binom{3000}{25}$, que son mucho más que 80. De manera que no se obtendrá una verdadera distribución muestral de las medias, sino únicamente una distribución muestral *experimental*. De cualquier manera, dado que el número de muestras es grande, habrá una estrecha coincidencia entre las dos distribuciones muestrales. Por lo tanto, la media y la desviación estándar esperadas serán muy semejantes a las de la distribución teórica. Se tiene:

a) $$\mu_{\bar{X}} = \mu = 68.0\,\text{in} \qquad \text{y} \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = \frac{3}{\sqrt{25}} = 0.6\,\text{in}$$

b) $$\mu_{\bar{X}} = 68.0\,\text{in} \qquad \text{y} \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}\sqrt{\frac{N_p - N}{N_p - 1}} = \frac{3}{\sqrt{25}}\sqrt{\frac{3\,000 - 25}{3\,000 - 1}}$$

que es apenas ligeramente menor a 0.6 in y por lo tanto, para fines prácticos, puede considerarse igual a la del muestreo con reposición.

De esta manera, se espera que la distribución muestral experimental de las medias esté distribuida de manera aproximadamente normal con media 68.0 in y desviación estándar 0.6 in.

8.4 ¿En cuántas de las muestras del problema 8.3 se esperaría encontrar que la media: *a*) estuviera entre 66.8 y 68.3 in y *b*) fuera menor a 66.4 in?

SOLUCIÓN

La media $\bar{X}$ de una muestra, en unidades estándar, está dada por

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - 68.0}{0.6}$$

a) $$66.8 \text{ en unidades estándar} = \frac{66.8 - 68.0}{0.6} = -2.0$$

$$68.3 \text{ en unidades estándar} = \frac{68.3 - 68.0}{0.6} = 0.5$$

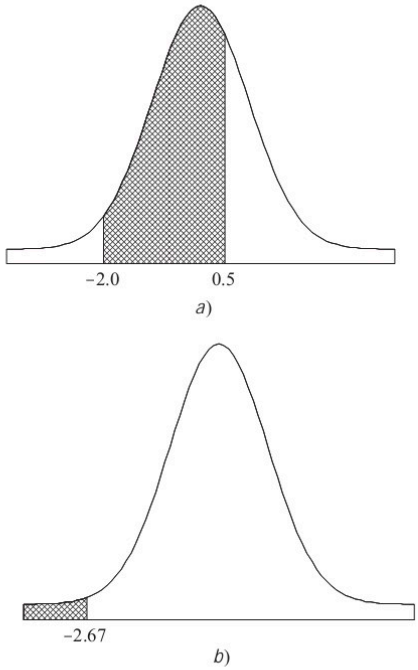


Figura 8-2 Áreas bajo la curva normal estándar. *a*) En esta curva normal estándar se muestra el área entre $z = -2$ y $z = 0.5$. *b*) En esta curva normal estándar se muestra el área a la izquierda de $z = -2.67$.

Como se muestra en la figura 8-2*a*).

La proporción de muestras cuya media está entre 66.8 y 68.3 in

$$= (\text{área bajo la curva normal entre } z = -2.0 \text{ y } z = 0.5)$$
$$= (\text{área entre } z = -2 \text{ y } z = 0) + (\text{área entre } z = 0 \text{ y } z = 0.5)$$
$$= 0.4772 + 0.1915 = 0.6687$$

Por lo tanto, la cantidad esperada de muestras es $(80)(0.6687) = 56.496$, o 53.

b)
$$66.4 \text{ en unidades estándar} = \frac{66.4 - 68.0}{0.6} = -2.67$$

Como se muestra en la figura 8-2*b*).

Proporción de las muestras que tienen una media

$$\text{menor que 66.4 in} = (\text{área bajo la curva normal a la izquierda de } z = -2.67)$$
$$= (\text{área a la izquierda de } z = 0)$$
$$- (\text{área entre } z = -2.67 \text{ y } z = 0)$$
$$= 0.5 - 0.4962 = 0.0038$$

Por lo tanto, el número de muestras esperada es $(80)(0.0038) = 0.304$ o cero.

8.5 Se tienen 500 balines cuyo peso medio es 5.02 gramos (g) y cuya desviación estándar es 0.30 g. Encontrar la probabilidad de que todos los balines de una muestra aleatoria de 100 balines, tomada de estos balines, pese: *a*) entre 496 y 500 g y *b*) más de 510 g.

SOLUCIÓN

Para la distribución muestral de las medias, $\mu_{\bar{X}} = \mu = 5.02\,\text{g}$, y

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}\sqrt{\frac{N_\text{p} - N}{N_\text{p} - 1}} = \frac{0.30}{\sqrt{100}}\sqrt{\frac{500 - 100}{500 - 1}} = 0.027\,\text{g}$$

a) El peso de los 100 balines, juntos, estará entre 496 y 500 g, si el peso medio de los balines se encuentra entre 4.96 y 5.00 g.

$$4.96 \text{ en unidades estándar} = \frac{4.96 - 5.02}{0.0027} = -2.22$$

$$5.00 \text{ en unidades estándar} = \frac{5.00 - 5.02}{0.027} = -0.74$$

Como se muestra en la figura 8-3*a*),

$$\begin{aligned}\text{Probabilidad buscada} &= (\text{área entre } z = -2.22 \text{ y } z = -0.74)\\ &= (\text{área entre } z = -2.22 \text{ y } z = 0) - (\text{área entre } z = -0.74 \text{ y } z = 0)\\ &= 0.4868 - 0.2704 = 0.2164\end{aligned}$$

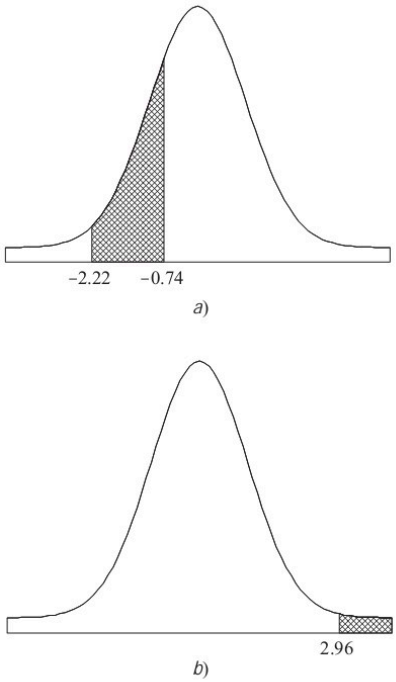


Figura 8-3 Las probabilidades muestrales se encuentran como áreas bajo la curva normal estándar. *a*) Curva normal estándar en la que se muestra el área entre $z = -2.22$ y $z = -0.74$; *b*) curva normal estándar en la que se muestra el área a la derecha de $z = 2.96$.

b) El peso de los 100 balines juntos será mayor a 510 g si el peso medio de los balines es mayor a 5.10 g.

$$5.10 \text{ en unidades estándar} = \frac{5.10 - 5.02}{0.027} = 2.96$$

Como se muestra la figura 8-3*b*),

$$\begin{aligned}\text{Probabilidad buscada} &= (\text{área a la derecha de } z = 2.96)\\ &= (\text{área a la derecha de } z = 0) - (\text{área entre } z = 0 \text{ y } z = 2.96)\\ &= 0.5 - 0.4985 = 0.0015\end{aligned}$$

Por lo tanto, sólo hay 3 posibilidades en 2 000 de obtener una muestra de 100 balines que juntos pesen más de 510 g.

8.6 *a*) Mostrar la manera de tomar, de la tabla 2.1, 30 muestras aleatorias (con reposición) de 4 estudiantes cada una, empleando números aleatorios.

b) Encontrar la media y la desviación estándar de la distribución muestral de las medias del inciso *a*).

c) Comparar los resultados del inciso *b*) con los valores teóricos y explicar cualquier discrepancia.

SOLUCIÓN

a) Para enumerar cada uno de los 100 estudiantes se emplean los dígitos: 00, 01, 02, . . . , 99 (ver la tabla 8.3). Por lo tanto, los 5 estudiantes cuya estatura está en el intervalo 60-62 están numerados 00-04; los dieciocho estudiantes cuya estatura está en el intervalo 63-65 están numerados 05-22, etc. Al número de cada estudiante se le llama *número de muestreo*.

Tabla 8.3

Estatura (in)	Frecuencias	Número de muestreo
60-62	5	00-04
63-65	18	05-22
66-68	42	23-64
69-71	27	65-91
72-74	8	92-99

Después, se extraen números de muestreo de una tabla de números aleatorios (apéndice IX). En el primer renglón se encuentra la secuencia 51, 77, 27, 46, 40, etc., la que será considerada como los números del muestreo aleatorios, cada uno de los cuales dará la estatura de determinado estudiante. Así, 51 corresponde a un estudiante cuya estatura está en el intervalo 66-68 in, estatura que se toma como 67 in (la marca de clase). De igual manera, 77, 27 y 46 dan las estaturas 70, 67 y 67 in, respectivamente.

Mediante este proceso se obtiene la tabla 8.4, en la que se muestra el número de muestreo extraído, la estatura correspondiente y la estatura promedio de cada una de las 30 muestras. Es necesario decir que aunque se han tomado los números aleatorios del primer renglón de la tabla, igualmente podría haberse partido de cualquier otro lugar de la tabla y elegir cualquier otro patrón específico.

b) En la tabla 8.5 se da la distribución de frecuencias de las medias muestrales de las estaturas obtenidas en el inciso *a*). Ésta es una *distribución muestral de las medias*. La media y la desviación estándar se obtienen, como de costumbre, empleando un método de compilación de los capítulos 3 y 4:

$$\text{Media} = A + c\bar{u} = A + \frac{c\sum fu}{N} = 67.00 + \frac{(0.75)(23)}{30} = 67.58\,\text{in}$$

$$\text{Desviación estándar} = c\sqrt{\overline{u^2} - \bar{u}^2} = c\sqrt{\frac{\sum fu^2}{N} - \left(\frac{\sum fu}{N}\right)^2} = 0.75\sqrt{\frac{123}{30} - \left(\frac{23}{30}\right)^2} = 1.41\,\text{in}$$

c) La media teórica de la distribución muestral de las medias, dada por $\mu_{\bar{X}}$, debe ser igual a la media poblacional μ, que es 67.45 in (ver problema 3.22), lo que coincide con el valor 67.58 del inciso *b*).

La desviación estándar teórica (error estándar) de la distribución muestral de las medias, dada por $\sigma_{\bar{X}}$, debe ser igual a $\sigma/\sqrt{N}$, donde la desviación estándar poblacional es $\sigma = 2.92$ in (ver problema 4.17) y el tamaño de la muestra es $N = 4$. Como $\sigma/\sqrt{N} = 2.92/\sqrt{4} = 1.46$ in, esto coincide con el valor 1.41 in del inciso *b*). Las ligeras discrepancias resultan de que sólo se tomaron 30 muestras y de que el tamaño de la muestra es pequeño.

Tabla 8.4

Números muestrales extraídos	Estaturas correspondientes	Estatura media	Números muestrales extraídos	Estaturas correspondientes	Estatura media
1. 51, 77, 27, 46	67, 70, 67, 67	67.75	**16.** 11, 64, 55, 58	64, 67, 67, 67	66.25
2. 40, 42, 33, 12	67, 67, 67, 64	66.25	**17.** 70, 56, 97, 43	70, 67, 73, 67	69.25
3. 90, 44, 46, 62	70, 67, 67, 67	67.75	**18.** 74, 28, 93, 50	70, 67, 73, 67	69.25
4. 16, 28, 98, 93	64, 67, 73, 73	69.25	**19.** 79, 42, 71, 30	70, 67, 70, 67	68.50
5. 58, 20, 41, 86	67, 64, 67, 70	67.00	**20.** 58, 60, 21, 33	67, 67, 64, 67	66.25
6. 19, 64, 08, 70	64, 67, 64, 70	66.25	**21.** 75, 79, 74, 54	70, 70, 70, 67	69.25
7. 56, 24, 03, 32	67, 67, 61, 67	65.50	**22.** 06, 31, 04, 18	64, 67, 61, 64	64.00
8. 34, 91, 83, 58	67, 70, 70, 67	68.50	**23.** 67, 07, 12, 97	70, 64, 64, 73	67.75
9. 70, 65, 68, 21	70, 70, 70, 64	68.50	**24.** 31, 71, 69, 88	67, 70, 70, 70	69.25
10. 96, 02, 13, 87	73, 61, 64, 70	67.00	**25.** 11, 64, 21, 87	64, 67, 64, 70	66.25
11. 76, 10, 51, 08	70, 64, 67, 64	66.25	**26.** 03, 58, 57, 93	61, 67, 67, 73	67.00
12. 63, 97, 45, 39	67, 73, 67, 67	68.50	**27.** 53, 81, 93, 88	67, 70, 73, 70	70.00
13. 05, 81, 45, 93	64, 70, 67, 73	68.50	**28.** 23, 22, 96, 79	67, 64, 73, 70	68.50
14. 96, 01, 73, 52	73, 61, 70, 67	67.75	**29.** 98, 56, 59, 36	73, 67, 67, 67	68.50
15. 07, 82, 54, 24	64, 70, 67, 67	67.00	**30.** 08, 15, 08, 84	64, 64, 64, 70	65.50

Tabla 8.5

Media muestral	Cuenta	f	u	fu	fu^2
64.00	/	1	−4	−4	16
64.75		0	−3	0	0
65.50	//	2	−2	−4	8
66.25	𝍸 /	6	−1	−6	6
$A \rightarrow$ 67.00	𝍸	4	0	0	0
67.75	𝍸	4	1	4	4
68.50	𝍸 //	7	2	14	28
69.25	𝍸	5	3	15	45
70.00	/	1	4	4	16
		$\sum f = N = 30$		$\sum fu = 23$	$\sum fu^2 = 123$

DISTRIBUCIÓN MUESTRAL DE PROPORCIONES

8.7 Encontrar la probabilidad de que en 120 lanzamientos de una moneda: *a*) menos del 40% o más del 60% sean cara y *b*) $\frac{5}{8}$ o más sean cara.

SOLUCIÓN

Primer método

Los 120 lanzamientos de la moneda se consideran una muestra de la población infinita de todos los posibles lanzamientos de una moneda. En esta población la probabilidad de obtener cara es $p = \frac{1}{2}$ y la probabilidad de obtener cruz es $q = 1 - p = \frac{1}{2}$.

a) La probabilidad que se busca es que en 120 lanzamientos, la cantidad de caras sea menor a 48 o mayor a 72. Se procederá, como en el capítulo 7, empleando la aproximación normal a la binomial. Como el número de caras es una variable discreta, se busca la probabilidad de que el número de caras sea menor a 47.5 o mayor a 72.5.

$$\mu = \text{número esperado de caras} = Np = 120(\tfrac{1}{2}) = 60 \qquad \text{y} \qquad \sigma = \sqrt{Npq} = \sqrt{(120)(\tfrac{1}{2})(\tfrac{1}{2})} = 5.48$$

$$47.5 \text{ en unidades estándar} = \frac{47.5 - 60}{5.48} = -2.28$$

$$72.5 \text{ en unidades estándar} = \frac{72.5 - 60}{5.48} = 2.28$$

Como se muestra en la figura 8-4,

$$\begin{aligned}\text{Probabilidad buscada} &= (\text{área a la izquierda de } -2.28 \text{ más área a la derecha de } 2.28)\\ &= (2(0.0113) = 0.0226)\end{aligned}$$

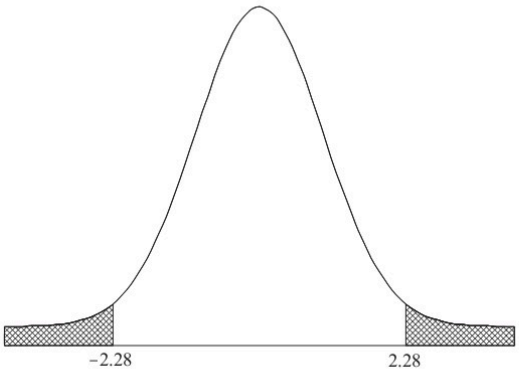


Figura 8-4 En la aproximación normal a la binomial se usa la curva normal estándar.

Segundo método

$$\mu_P = p = \tfrac{1}{2} = 0.50 \qquad \sigma_P = \sqrt{\frac{pq}{N}} = \sqrt{\frac{(\frac{1}{2})(\frac{1}{2})}{120}} = 0.0456$$

$$40\% \text{ en unidades estándar} = \frac{0.40 - 0.50}{0.0456} = -2.19$$

$$60\% \text{ en unidades estándar} = \frac{0.60 - 0.50}{0.0456} = 2.19$$

$$\begin{aligned}\text{Probabilidad buscada} &= (\text{área a la izquierda de } -2.19 \text{ más área a la derecha de } 2.19)\\ &= (2(0.0143) = 0.0286)\end{aligned}$$

Aunque este resultado es exacto a dos cifras significativas, no hay una coincidencia exacta debido a que no se usó el hecho de que una proporción es en realidad una variable discreta. Para tomar en cuenta esto, a 0.40 se le resta

$1/2N = 1/2(120)$ y a 0.60 se le suma $1/2N = 1/2(120)$; como $1/240 = 0.00417$, las proporciones buscadas son, en unidades estándar,

$$\frac{0.40 - 0.00417 - 0.50}{0.0456} = -2.28 \qquad \text{y} \qquad \frac{0.60 + 0.00417 - 0.50}{0.0456} = 2.28$$

con lo que se obtiene coincidencia con el primer método.

Obsérvese que $(0.40 - 0.00417)$ y $(0.60 + 0.00417)$ corresponden a las proporciones 47.5/120 y 72.5/120 usadas en el primer método.

b) Usando el segundo método del inciso *a*) se encuentra que como $\frac{5}{8} = 0.6250$,

$$(0.6250 - 0.00417) \text{ en unidades estándar} = \frac{0.6250 - 0.00417 - 0.50}{0.0456} = 2.65$$

$$\begin{aligned} \text{Probabilidad buscada} &= (\text{área bajo la curva normal a la derecha de } z = 2.65) \\ &= (\text{área a la derecha de } z = 0) - (\text{área entre } z = 0 \text{ y } z = 2.65) \\ &= 0.5 - 0.4960 = 0.0040 \end{aligned}$$

8.8 Cada una de las 500 personas de un grupo lanza una moneda 120 veces. ¿Cuántas personas pueden esperar: *a*) que entre el 40% y el 60% de sus lanzamientos sean cara y *b*) en $\frac{5}{8}$, o más, de sus lanzamientos obtener cara?

SOLUCIÓN

Este problema está estrechamente relacionado con el problema 8.7. Aquí se consideran 500 muestras, cada una de tamaño 120, tomadas de la población infinita de todos los posibles lanzamientos de una moneda.

a) En el inciso *a*) del problema 8.7 se establece que de todas las muestras posibles, cada una consistente en 120 lanzamientos de una moneda, se puede esperar que en el 97.74% de las mismas se tenga entre 40% y 60% de caras. Entonces, en 500 muestras se puede esperar que aproximadamente (97.74% de 500) = 489 muestras tengan esta propiedad. Se concluye que alrededor de 489 personas pueden esperar que en su experimento entre 40% y 60% sean caras.

Es interesante observar que hay $500 - 489 = 11$ personas para las que se espera que el porcentaje de caras que obtengan no esté entre el 40 y el 60%. Estas personas pueden concluir, con razón, que su moneda esté cargada. Este tipo de error es un *riesgo* siempre presente cuando se trata con probabilidad.

b) Razonando como en el inciso *a*) se concluye que aproximadamente $(500)(0.0040) = 2$ personas obtendrán caras en $\frac{5}{8}$, o más, de sus lanzamientos.

8.9 Se encuentra que el 2% de las herramientas producidas con determinada máquina están defectuosas. ¿Cuál es la probabilidad de que en un pedido de 400 de estas herramientas: *a*) 3% o más y *b*) 2% o menos resulten defectuosas?

SOLUCIÓN

$$\mu_P = p = 0.02 \qquad \text{y} \qquad \sigma_P = \sqrt{\frac{pq}{N}} = \sqrt{\frac{(0.02)(0.98)}{400}} = \frac{0.14}{20} = 0.007$$

a) **Primer método**

Empleando la corrección para variables discretas, $1/2N = 1/800 = 0.00125$, se tiene

$$(0.03 - 0.00125) \text{ en unidades estándar} = \frac{0.03 - 0.00125 - 0.02}{0.007} = 1.25$$

$$\text{Probabilidad buscada} = (\text{área bajo la curva normal a la derecha de } z = 1.25) = 0.1056$$

Si no se usa la corrección, se obtiene 0.0764.

Otro método

(3% de 400) = 12 herramientas defectuosas. Considerando la variable como una variable continua, 12 o más herramientas significa 11.5 o más.

$$\bar{X} = (2\% \text{ de } 400) = 8 \qquad \text{y} \qquad \sigma = \sqrt{Npq} = \sqrt{(400)(0.02)(0.98)} = 2.8$$

Entonces, 11.5 en unidades estándar = (11.5 – 8)/2.8 = 1.25, y como se encontró antes, la probabilidad buscada es 0.1056.

b) $$(0.02 + 0.00125) \text{ en unidades estándar} = \frac{0.02 + 0.00125 - 0.02}{0.007} = 0.18$$

$$\text{Probabilidad buscada} = (\text{área bajo la curva normal a la izquierda de } z = 0.18)$$
$$= 0.5000 + 0.0714 = 0.5714$$

Si no se usa la corrección, se obtiene 0.5000. También se puede usar el segundo método del inciso *a*).

8.10 Como resultado de una elección se observa que determinado candidato obtuvo 46% de los votos. Determinar la probabilidad de que en una encuesta realizada a: *a*) 200 y *b*) 1 000 personas elegidas al azar de la población de votantes se hubiera obtenido una mayoría de votos a favor de este candidato.

SOLUCIÓN

a) $$\mu_P = p = 0.46 \qquad \text{y} \qquad \sigma_P = \sqrt{\frac{pq}{N}} = \sqrt{\frac{(0.46)(0.54)}{200}} = 0.0352$$

Como $1/2N = 1/400 = 0.0025$, se tiene una mayoría en la muestra si la proporción a favor de este candidato es 0.50 + 0.0025 = 0.5025 o más. (Esta proporción puede obtenerse también observando que una mayoría es 101 o más, pero considerada como variable continua esto corresponde a 100.5, con lo que la proporción es 100.5/200 = 0.5025.)

$$0.5025 \text{ en unidades estándar} = \frac{0.5025 - 0.46}{0.0352} = 1.21$$

$$\text{Probabilidad buscada} = (\text{área bajo la curva normal a la derecha de } z = 1.21)$$
$$= 0.5000 - 0.3869 = 0.1131$$

b) $$\mu_p = p = 0.46 \qquad \text{y} \qquad \sigma_P = \sqrt{\frac{pq}{N}} = \sqrt{\frac{(0.46)(0.54)}{1\,000}} = 0.0158$$

$$0.5025 \text{ en unidades estándar} = \frac{0.5025 - 0.46}{0.0158} = 2.69$$

$$\text{Probabilidad buscada} = (\text{área bajo la curva normal a la derecha de } z = 2.69)$$
$$= 0.5000 - 0.4964 = 0.0036$$

DISTRIBUCIÓN MUESTRAL DE DIFERENCIAS Y DE SUMAS

8.11 Sea U_1 una variable que representa los elementos de la población 3, 7, 8 y U_2 una variable que representa los elementos de la población 2, 4. Calcular: *a*) μ_{U1}, *b*) μ_{U2}, *c*) μ_{U1-U2}, *d*) σ_{U1}, *e*) σ_{U2} y *f*) σ_{U1-U2}.

SOLUCIÓN

a) μ_{U1} = media de la población $U_1 = \frac{1}{3}(3 + 7 + 8) = 6$

b) μ_{U2} = media de la población $U_2 = \frac{1}{2}(2 + 4) = 3$

c) Esta población consta de todas las diferencias entre los miembros de U_1 y U_2, que es

$$\begin{matrix} 3-2 & 7-2 & 8-2 \\ 3-4 & 7-4 & 8-4 \end{matrix} \quad \text{o} \quad \begin{matrix} 1 & 5 & 6 \\ -1 & 3 & 4 \end{matrix}$$

Por lo tanto, $$\mu_{U1-U2} = \text{media de } (U_1 - U_2) = \frac{1+5+6+(-1)+3+4}{6} = 3$$

Esto ilustra que $\mu_{U1-U2} = \mu_{U1} - \mu_{U2}$, como se ve en los incisos *a*) y *b*).

d) $$\sigma_{U1}^2 = \text{varianza de la población } U_1 = \frac{(3-6)^2 + (7-6)^2 + (8-6)^2}{3} = \frac{14}{3}$$

o bien $$\sigma_{U1} = \sqrt{\frac{14}{3}}$$

e) $$\sigma_{U2}^2 = \text{varianza de la población } U_2 = \frac{(2-3)^2 + (4-3)^2}{2} = 1 \qquad \text{o bien} \qquad \sigma_{U2} = 1$$

f) $$\sigma_{U1-U2}^2 = \text{varianza de la población } (U_1 - U_2)$$

$$= \frac{(1-3)^2 + (5-3)^2 + (6-3)^2 + (-1-3)^2 + (3-3)^2 + (4-3)^2}{6} = \frac{17}{3}$$

o $$\sigma_{U1-U2} = \sqrt{\frac{17}{3}}$$

Esto ilustra que para muestras independientes $\sigma_{U1-U2} = \sqrt{\sigma_{U1}^2 + \sigma_{U2}^2}$, como se ve en los incisos *d*) y *e*).

8.12 El tiempo medio de vida de los focos del fabricante *A* es 1 400 horas (h) y su desviación estándar es 200 h, en tanto que el tiempo medio de vida de los focos del fabricante *B* es 1 200 h y su desviación estándar es 100 h. Si se prueban muestras aleatorias de 125 focos de cada fabricante, ¿cuál es la probabilidad de que el tiempo medio de vida de los focos del fabricante *A* sea por lo menos: *a*) 160 h y *b*) 250 h mayor que el del fabricante *B*?

SOLUCIÓN

Sean $\bar{X}_A$ y $\bar{X}_B$ los tiempos medios de vida en las muestras de *A* y de *B*, respectivamente. Entonces

$$\mu_{\bar{X}_A - \bar{X}_B} = \mu_{\bar{X}_A} - \mu_{\bar{X}_B} = 1\,400 - 1\,200 = 200 \text{ h}$$

y $$\sigma_{\bar{X}_A - \bar{X}_B} = \sqrt{\frac{\sigma_A^2}{N_A} + \frac{\sigma_B^2}{N_B}} = \sqrt{\frac{(100)^2}{125} + \frac{(200)^2}{125}} = 20 \text{ h}$$

La variable estandarizada que corresponde a la diferencia entre las medias es

$$z = \frac{(\bar{X}_A - \bar{X}_B) - (\mu_{\bar{X}_A - \bar{X}_B})}{\sigma_{\bar{X}_A - \bar{X}_B}} = \frac{(\bar{X}_A - \bar{X}_B) - 200}{20}$$

que está distribuida casi normalmente.

a) La diferencia de 160 h en unidades estándar es $(160 - 200)/20 = -2$. Por lo tanto

$$\text{Probabilidad buscada} = (\text{área bajo la curva normal a la derecha de } z = -2)$$
$$= 0.5000 + 0.4772 = 0.9772$$

b) La diferencia de 250 h en unidades estándar es (250 – 200)/20 = 2.50. Por lo tanto

$$\text{Probabilidad buscada} = (\text{área bajo la curva normal a la derecha de } z = 2.50)$$
$$= 0.5000 - 0.4938 = 0.0062$$

8.13 Los balines de determinada marca tienen un peso promedio de 0.50 g y su desviación estándar es 0.02 g. ¿Cuál es la probabilidad de que entre dos lotes, cada uno de 1 000 balines, haya una diferencia de peso de más de 2 g?

SOLUCIÓN

Sean $\bar{X}_1$ y $\bar{X}_2$ las medias de los pesos de los balines de los dos lotes. Entonces

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = 0.50 - 0.50 = 0$$

y

$$\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{(0.02)^2}{1\,000} + \frac{(0.02)^2}{1\,000}} = 0.000895$$

La variable estandarizada correspondiente a la diferencia entre las medias es

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{0.000895}$$

que está distribuida en forma aproximadamente normal.

Una diferencia de 2 g entre los lotes es equivalente a una diferencia de $2/1\,000 = 0.002$ g entre las medias. Esto puede ocurrir si $\bar{X}_1 - \bar{X}_2 \geq 0.002$ o si $\bar{X}_1 - X_2 \leq -0.002$; es decir,

$$z \geq \frac{0.002 - 0}{0.000895} = 2.23 \qquad \text{o} \qquad z \leq \frac{-0.002 - 0}{0.000895} = -2.23$$

Entonces $\Pr\{z \geq 2.23 \text{ o } z \leq -2.23\} = \Pr\{z \geq 2.23\} + \Pr\{z \leq -2.23\} = 2(0.5000 - 0.4871) = 0.0258$.

8.14 *A* y *B* juegan un partido que consiste en que cada uno lance 50 monedas. *A* gana el partido si obtiene 5 o más caras que *B*; si no, *B* lo gana. Determinar las posibilidades en contra de que *A* gane un juego.

SOLUCIÓN

Sean P_A y P_B las proporciones de caras obtenidas por *A* y por *B*, respectivamente. Si se supone que las monedas no están cargadas, la probabilidad p de obtener una cara es $\frac{1}{2}$. Entonces

$$\mu_{P_A-P_B} = \mu_{P_A} - \mu_{P_B} = 0$$

y

$$\sigma_{P_A-P_B} = \sqrt{\sigma_{P_A}^2 + \sigma_{P_B}^2} = \sqrt{\frac{pq}{N_A} + \frac{pq}{N_B}} = \sqrt{\frac{2(\frac{1}{2})(\frac{1}{2})}{50}} = 0.10$$

La variable estandarizada correspondiente a esta diferencia entre las proporciones es $z = (P_A - P_B - 0)/0.10$.

Considerando la variable como variable continua, 5 o más caras corresponde a 4.5 o más caras, de manera que la diferencia entre las proporciones debe ser de $4.5/50 = 0.09$ o más; es decir, z es mayor o igual a $(0.09 - 0)/0.10 = 0.9$ (o bien $z \geq 0.9$). La probabilidad de esto es el área bajo la curva normal a la derecha de $z = 0.9$, la cual es $(0.5000 - 0.3159) = 0.1841$.

De manera que las posibilidades en contra de que *A* gane son $(1 - 0.1841):0.1841 = 0.8159:0.1841$, o 4.43 a 1.

8.15 Dos distancias se miden como 27.3 centímetros (cm) y 15.6 cm con desviaciones estándar (errores estándar) de 0.16 y 0.08 cm, respectivamente. Determinar la media y la desviación estándar de: *a*) la suma y *b*) la diferencia de estas distancias.

SOLUCIÓN

Si las distancias se denotan D_1 y D_2, entonces:

a)

$$\mu_{D1+D2} = \mu_{D1} + \mu_{D2} = 27.3 + 15.6 = 42.9\,\text{cm}$$

$$\sigma_{D1+D2} = \sqrt{\sigma_{D1}^2 + \sigma_{D2}^2} = \sqrt{(0.16)^2 + (0.08)^2} = 0.18\,\text{cm}$$

b)

$$\mu_{D1-D2} = \mu_{D1} - \mu_{D2} = 27.3 - 15.6 = 11.7\,\text{cm}$$

$$\sigma_{D1-D2} = \sqrt{\sigma_{D1}^2 + \sigma_{D2}^2} = \sqrt{(0.16)^2 + (0.08)^2} = 0.18\,\text{cm}$$

8.16 La media del tiempo de vida de determinado tipo de foco es 1 500 h y la desviación estándar es 150 h. Se conectan tres de estos focos de manera que cuando uno se funda, otro empiece a funcionar. Suponiendo que los tiempos de vida estén distribuidos normalmente, ¿cuál es la probabilidad de que la iluminación dure: *a*) por lo menos 5 000 h y *b*) a lo mucho 4 200 h?

SOLUCIÓN

Suponga que los tiempos de vida son L_1, L_2 y L_3. Entonces

$$\mu_{L1+L2+L3} = \mu_{L1} + \mu_{L2} + \mu_{L3} = 1\,500 + 1\,500 + 1\,500 = 4\,500 \text{ h}$$

$$\sigma_{L1+L2+L3} = \sqrt{\sigma_{L1}^2 + \sigma_{L2}^2 + \sigma_{L3}^2} = \sqrt{3(150)^2} = 260 \text{ h}$$

a)
$$5\,000 \text{ h en unidades estándar} = \frac{5\,000 - 4\,500}{260} = 1.92$$

$$\text{Probabilidad buscada} = (\text{área bajo la curva normal a la derecha de } z = 1.92)$$
$$= 0.5000 - 0.4726 = 0.0274$$

b)
$$4\,200 \text{ h en unidades estándar} = \frac{4\,200 - 4\,500}{260} = -1.15$$

$$\text{Probabilidad buscada} = (\text{área bajo la curva normal a la izquierda de } z = -1.15)$$
$$= 0.5000 - 0.3749 = 0.1251$$

DEMOSTRACIONES DE LA TEORÍA ELEMENTAL DEL MUESTREO EMPLEANDO SOFTWARE

8.17 En una universidad, 1/3 de los estudiantes toma 9 horas de crédito, 1/3 toma 12 horas de crédito y 1/3 toma 15 horas de crédito. Si X representa las horas de crédito que toma un estudiante, la distribución de X es $p(x) = 1/3$ para $x = 9$, 12 y 15. Encontrar la media y la varianza de X. ¿Qué tipo de distribución tiene X?

SOLUCIÓN

La media de X es $\mu = \sum xp(x) = 9(1/3) + 12(1/3) + 15(1/3) = 12$. La varianza de X es $\sigma^2 = \sum x^2 p(x) - \mu^2 = 81(1/3) + 144(1/3) + 225(1/3) - 144 = 150 - 144 = 6$. La distribución de X es uniforme.

8.18 Enumerar todas las muestras de tamaño $n = 2$ que pueden tomarse (con reposición) de la población del problema 8.17. Usar el asistente para gráficos de EXCEL para graficar la distribución muestral de la media y mostrar que $\mu_{\bar{x}} = \mu$ y que $\sigma_{\bar{x}}^2 = \sigma^2/2$.

SOLUCIÓN

A	B	C	D	E	F	G
		media	x barra	p(x barra)	x barra × p(x barra)	x barra^2 × p(x barra)
9	9	9	9	0.111111	1	9
9	12	10.5	10.5	0.222222	2.333333333	24.5
9	15	12	12	0.333333	4	48
12	9	10.5	13.5	0.222222	3	40.5
12	12	12	15	0.111111	1.666666667	25
12	15	13.5			12	147
15	9	12				
15	12	13.5				
15	15	15				

La hoja de cálculo de EXCEL muestra en A y en B los valores muestrales posibles y en C las medias. La distribución muestral de x barra se construye y se da en D y E. En C2 se ingresa la función =AVERAGE(A2:B2), se hace clic y se arrastra de C2 a C10. Como la población es uniforme, cada muestra tiene probabilidad 1/9 de ser elegida. La media muestral se representa por *x* barra. La media de las medias muestrales es $\mu_{\bar{x}} = \Sigma \bar{x} p(\bar{x})$ y se calcula de F2 a F6. La función =SUM(F2:F6) se ingresa en F7 y se obtiene 12, mostrando que $\mu_{\bar{x}} = \mu$. La varianza de las medias muestrales es $\sigma_{\bar{x}}^2 = \Sigma \bar{x}^2 p(\bar{x}) - \mu_{\bar{x}}^2$ y se calcula como sigue. De G2 a G6 se calcula $\Sigma \bar{x}^2 p(\bar{x})$. En G7 se ingresa la función =SUM(G2:G6), que es igual a 147. Restando 12^2 o bien144 de 147, se obtiene 3, que es $\sigma_{\bar{x}}^2 = \sigma^2/2$. En la figura 8-5 se muestra que con un tamaño de muestra 2, la distribución muestral de $\bar{x}$ es un poco parecida a una distribución normal. Las probabilidades mayores se encuentran cerca de 12 y éstas disminuyen hacia la derecha y hacia la izquierda de 12.

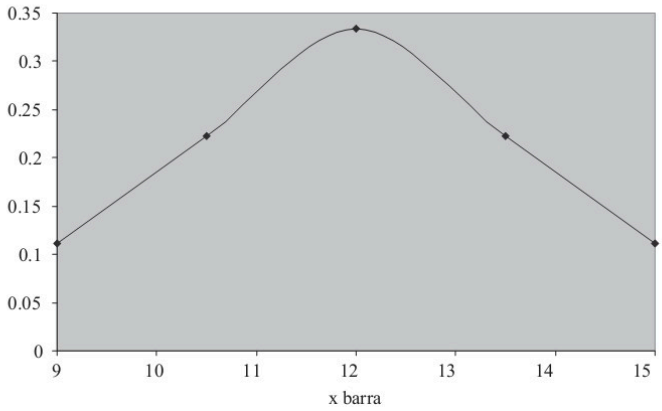


Figura 8-5 Distribución de x barra para *n* = 2.

8.19 Enlistar todas las muestras de tamaño $n = 3$ que se pueden obtener con reposición de la población del problema 8.17. Usar EXCEL para construir la distribución muestral de la media. Para graficar la distribución muestral de la media se usa el asistente para gráficos de EXCEL. Mostrar que $\mu_{\bar{x}} = \mu$ y que $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{3}$.

SOLUCIÓN

A	B	C	D	E	F	G	H
			media	x barra	p(x barra)	x barra*p(x barra)	x barra^2p(x barra)
9	9	9	9	9	0.037037037	0.333333333	3
9	9	12	10	10	0.111111111	1.111111111	11.11111111
9	9	15	11	11	0.222222222	2.444444444	26.88888889
9	12	9	10	12	0.259259259	3.111111111	37.33333333
9	12	12	11	13	0.222222222	2.888888889	37.55555556
9	12	15	12	14	0.111111111	1.555555556	21.77777778
9	15	9	11	15	0.037037037	0.555555556	8.333333333
9	15	12	12			12	146
9	15	15	13				
12	9	9	10				

Continuación

A	B	C	D	E	F	G	H
			media	x barra	p(x barra)	x barra*p(x barra)	x barra^2p(x barra)
12	9	12	11				
12	9	15	12				
12	12	9	11				
12	12	12	12				
12	12	15	13				
12	15	9	12				
12	15	12	13				
12	15	15	14				
15	9	9	11				
15	9	12	12				
15	9	15	13				
15	12	9	12				
15	12	12	13				
15	12	15	14				
15	15	9	13				
15	15	12	14				
15	15	15	15				

En la hoja de cálculo de EXCEL se muestran en A, B y C todos los valores muestrales, las medias se dan en D y la distribución muestral de x barra se calcula y se da en E y F. En D2 se ingresa la función =AVERAGE(A2:C2), se hace clic y se arrastra desde D2 hasta D28. Como esta población es uniforme, cada muestra tiene la probabilidad $1/27$ de ser elegida. La media muestral se representa por x barra. La media de las medias muestral es $\mu_{\bar{x}} = \Sigma \bar{x} p(\bar{x})$ y se calcula desde G2 hasta G8. La función =SUM(G2:G8) se ingresa en G9 y da como resultado 12, lo que demuestra que con muestras de tamaño $n = 3$ $\mu_{\bar{x}} = \mu$. La varianza de las medias muestrales es $\sigma_{\bar{x}}^2 = \Sigma \bar{x}^2 p(\bar{x}) - \mu_{\bar{x}}^2$ y se calcula como sigue. Desde H2 hasta H8 se calcula $\Sigma \bar{x}^2 p(\bar{x})$. En H9 se ingresa la función =SUM(H2:H8), que da como resultado 146. Restando 12^2, que es 144, de 146, se obtiene 2. Obsérvese que $\sigma_{\bar{x}}^2 = \sigma^2/3$. En la figura 8-6 se observa la tendencia de la distribución x barra a una distribución normal.

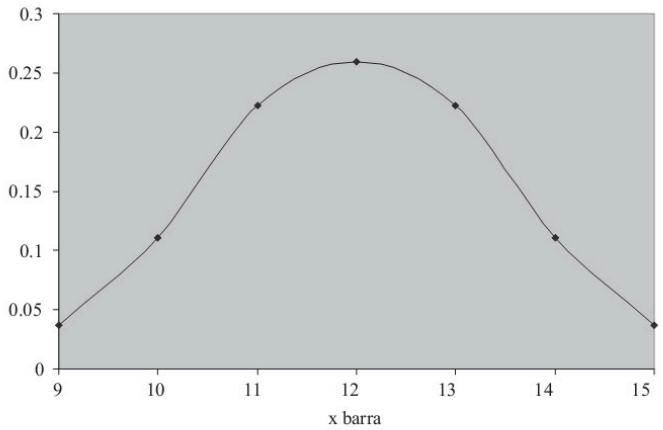


Figura 8-6 Distribución de x barra para $n = 3$.

8.20 Enlistar las 81 muestras de tamaño $n = 4$ (con reposición) que se pueden obtener de la población del problema 8.17. Usar EXCEL para construir la distribución muestral de las medias. Con el asistente para gráficos de EXCEL, graficar la distribución muestral de las medias, y mostrar que $\mu_{\bar{x}} = \mu$ y que $\sigma_{\bar{x}}^2 = \sigma^2/4$.

SOLUCIÓN

El método empleado en los problemas 8.18 y 8.19 se extiende a muestras de tamaño 4. En la hoja de cálculo de EXCEL se obtiene la siguiente distribución para x barra. Además, se puede demostrar que $\mu_{\bar{x}} = \mu$ y que $\sigma_{\bar{x}}^2 = \sigma^2/4$.

x barra	p(x barra)	x barra*p(x barra)	x barra^2p(x barra)
9	0.012345679	0.111111111	1
9.75	0.049382716	0.481481481	4.694444444
10.5	0.12345679	1.296296296	13.61111111
11.25	0.197530864	2.222222222	25
12	0.234567901	2.814814815	33.77777778
12.75	0.197530864	2.518518519	32.11111111
13.5	0.12345679	1.666666667	22.5
14.25	0.049382716	0.703703704	10.02777778
15	0.012345679	0.185185185	2.777777778
	1	12	145.5

En la figura 8-7 se muestra la gráfica de EXCEL de la distribución de x barra para muestras de tamaño 4.

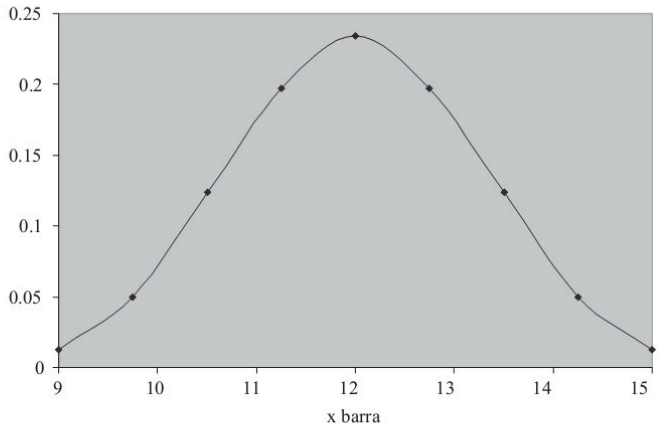


Figura 8-7 Distribución de x barra para muestras de $n = 4$.

PROBLEMAS SUPLEMENTARIOS

DISTRIBUCIÓN MUESTRAL DE LAS MEDIAS

8.21 Una población consta de los cuatro números 3, 7, 11 y 15. Considerar todas las posibles muestras con reposición de tamaño 2 que pueden obtenerse de esta población. Encontrar: *a*) la media poblacional, *b*) la desviación estándar poblacional,

c) la media de la distribución muestral de las medias y *d*) la desviación estándar de la distribución muestral de las medias. Verificar los incisos *c*) y *d*) directamente a partir de los incisos *a*) y *b*) empleando las fórmulas adecuadas.

8.22 Resolver el problema 8.21 si el muestreo se hace sin reposición.

8.23 Las masas de 1 500 balines están distribuidas de manera normal, siendo su media 22.40 g y su desviación estándar 0.048 g. Si de esta población se toman 300 muestras aleatorias de tamaño 36, determinar la media y la desviación estándar esperadas en la distribución muestral de las medias si el muestreo se hace: *a*) con reposición y *b*) sin reposición.

8.24 Resolver el problema 8.23 si la población consta de 72 balines.

8.25 ¿En cuántas de las muestras aleatorias del problema 8.23 la media: *a*) estará entre 22.39 y 22.41 g, *b*) será mayor a 22.42 g, *c*) será menor a 22.37 g y *d*) será menor a 22.38 g o mayor a 22.41 g?

8.26 La media de la vida útil de ciertos cinescopios fabricados por una empresa es 800 h y la desviación estándar es 60 h. Encontrar la probabilidad de que en una muestra aleatoria de 16 cinescopios la media del tiempo de vida: *a*) esté entre 790 y 810 h, *b*) sea menor a 785 h, *c*) sea mayor a 820 h y *d*) esté entre 770 y 830 h.

8.27 Repetir el problema 8.26 con una muestra aleatoria de 64 cinescopios. Explicar la diferencia.

8.28 Los paquetes que se reciben en una tienda departamental pesan en promedio 300 libras (lb) y su desviación estándar es de 50 lb. ¿Cuál es la probabilidad de que 25 paquetes recibidos al azar pesen más del límite de seguridad especificado en el elevador, que es 8 200 lb?

NÚMEROS ALEATORIOS

8.29 Repetir el problema 8.6 usando un conjunto diferente de números aleatorios y seleccionando: *a*) 15, *b*) 30, *c*) 45 y *d*) 60 muestras, con reposición, de tamaño 4. En cada caso, comparar con los resultados teóricos.

8.30 Repetir el problema 8.29 tomando muestras de tamaño: *a*) 2 y *b*) 8 con reposición, en lugar de tamaño 4 con reposición.

8.31 Repetir el problema 8.6, pero muestreando sin reposición. Comparar con los resultados teóricos.

8.32 *a*) Mostrar cómo se toman 30 muestras de tamaño 2 de la distribución del problema 3.61.

b) Calcular la media y la desviación estándar de la distribución muestral de las medias obtenida y compararla con los resultados teóricos.

8.33 Repetir el problema 8.32 empleando muestras de tamaño 4.

DISTRIBUCIÓN MUESTRAL DE PROPORCIONES

8.34 Encontrar la probabilidad de que de los 200 próximos niños que nazcan, *a*) menos de 40% sean varones, *b*) entre 43 y 57% sean niñas y *c*) más de 54% sean varones. Supóngase que existe la misma probabilidad de nacimiento de un niño que de una niña.

8.35 De 1 000 muestras, cada una de 200 niños, ¿en cuántas puede esperarse encontrar que: *a*) menos del 40% sean niños, *b*) entre 40 y 60% sean niñas y *c*) 53% o más sean niñas?

8.36 Repetir el problema 8.34 si las muestras son de 100 y no de 200 niños y explicar las diferencias resultantes.

8.37 Una urna contiene 80 canicas, de las cuales el 60% son rojas y el 40% son blancas. De 50 muestras, cada una de 20 canicas, tomadas de la urna con reposición, ¿en cuántas muestras se puede esperar que: *a*) haya el mismo número de canicas rojas que de canicas blancas, *b*) haya 12 canicas rojas y 8 canicas blancas, *c*) haya 8 canicas rojas y 12 canicas blancas y *d*) 10 o más canicas sean blancas?

8.38 Diseñar un experimento que tenga por objeto ilustrar los resultados del problema 8.37. En lugar de canicas rojas y blancas se pueden usar tiras de papel en las que se escriba R y B en las proporciones adecuadas. ¿Qué error se introduciría al usar dos conjuntos diferentes de monedas?

8.39 Un fabricante envía 1 000 lotes, cada uno de 100 bulbos eléctricos. Si es normal que el 5% de los bulbos esté defectuoso, ¿en cuántos de los lotes se esperaría: *a*) menos de 90 bulbos buenos y *b*) 98 o más bulbos buenos?

DISTRIBUCIONES MUESTRALES DE DIFERENCIA Y DE SUMAS

8.40 *A* y *B* fabrican cables que tienen una resistencia media a la ruptura de 4 000 lb y 4 500 lb, y desviaciones estándar de 300 lb y 200 lb, respectivamente. Si se prueban 100 cables del fabricante *A* y 50 cables del fabricante *B*, ¿cuál es la probabilidad de que la resistencia media a la ruptura de *B* sea: *a*) por lo menos 600 lb mayor que la de *A* y *b*) por lo menos 450 lb mayor que la de *A*?

8.41 En el problema 8.40, ¿cuáles son las probabilidades si se prueban 100 cables de cada fabricante? Explicar cualquier diferencia.

8.42 La puntuación media obtenida por los estudiantes en una prueba de aptitud es 72 puntos y la desviación estándar es 8 puntos. ¿Cuál es la probabilidad de que dos grupos de estudiantes, uno de 28 y otro de 36 estudiantes, difieran en la media de sus puntuaciones en: *a*) 3 o más puntos, *b*) 6 o más puntos y *c*) entre 2 y 5 puntos?

8.43 Una urna contiene 60 canicas rojas y 40 canicas blancas. De esta urna se extraen, con reposición, dos conjuntos de 30 canicas cada uno, y se van anotando sus colores. ¿Cuál es la probabilidad de que los dos conjuntos difieran en 8 o más canicas rojas?

8.44 Resolver el problema 8.43 si para obtener los dos conjuntos de canicas el muestreo se hace sin reposición.

8.45 Los resultados de una elección indican que un candidato obtuvo el 65% de los votos. Encontrar la probabilidad de que dos muestras aleatorias, cada una de 200 votantes, indiquen una diferencia mayor al 10% entre las proporciones de quienes votaron por el candidato.

8.46 Si U_1 y U_2 son los conjuntos de números del problema 8.11, verificar que: *a*) $\mu_{U1+U2} = \mu_{U1} + \mu_{U2}$ y *b*) $\sigma_{U1+U2} = \sqrt{\sigma_{U1}^2 + \sigma_{U2}^2}$.

8.47 Los valores que se obtienen al medir tres masas son 20.48, 35.97 y 62.34 g, con desviaciones estándar de 0.21, 0.46 y 0.54 g, respectivamente. Encontrar: *a*) la media y *b*) la desviación estándar de la suma de las masas.

8.48 El voltaje medio de una batería es 15.0 volts (V) y la desviación estándar es 0.2 V. ¿Cuál es la probabilidad de que cuatro de estas baterías conectadas en serie tengan, juntas, un voltaje de 60.8 V o más?

DEMOSTRACIONES DE LA TEORÍA ELEMENTAL DEL MUESTREO EMPLEANDO SOFTWARE

8.49 En una universidad la distribución de las horas crédito es como sigue:

x	6	9	12	15	18
$p(x)$	0.1	0.2	0.4	0.2	0.1

Encontrar μ y σ^2. Dar las 25 muestras (con reposición) de tamaño 2 que se pueden obtener, su media y sus probabilidades.

8.50 Graficar la distribución de probabilidad de x barra del problema 8.49, para $n = 2$.

8.51 Con los datos del problema 8.50, mostrar que $\mu_{\bar{x}} = \mu$ y $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{2}$.

8.52 Con los datos del problema 8.49, proporcionar y graficar la distribución de probabilidad de x barra para $n = 3$.

TEORÍA DE LA ESTIMACIÓN ESTADÍSTICA

9

ESTIMACIÓN DE PARÁMETROS

En el capítulo 8 se vio cómo emplear la teoría del muestreo para obtener información acerca de muestras extraídas en forma aleatoria de una población desconocida. Sin embargo, desde el punto de vista práctico, suele ser más importante poder inferir información acerca de una población a partir de muestras obtenidas de ella. De estos problemas se ocupa la *inferencia estadística* en la que se usan los principios de la teoría del muestreo.

Un problema importante de la inferencia estadística es la estimación de *parámetros poblacionales*, o simplemente *parámetros* (como, por ejemplo, la media y la varianza poblacionales), a partir de los correspondientes *estadísticos muestrales,* o simplemente *estadísticos* (por ejemplo, la media y la varianza muestrales). En este capítulo se analiza este problema.

ESTIMACIONES INSESGADAS

Si la media de la distribución muestral de un estadístico es igual al parámetro poblacional correspondiente se dice que el estadístico es un *estimador insesgado* del parámetro; si no es así, se dice que es un *estimador sesgado.* A los valores de estos estadísticos se les llama estimaciones *insesgadas* o *sesgadas*, respectivamente.

EJEMPLO 1 La media de la distribución muestral de las medias $\mu_{\bar{X}}$ es μ, la media poblacional. Por lo tanto, la media muestral $\bar{X}$ es una estimación insesgada de la media poblacional μ.

EJEMPLO 2 La media de la distribución muestral de las varianzas es

$$\mu_{s^2} = \frac{N-1}{N}\sigma^2$$

donde σ^2 es la varianza poblacional y N es el tamaño de la muestra (ver tabla 8.1). Por lo tanto, la varianza muestral s^2 es una estimación sesgada de la varianza poblacional σ^2. Empleando la varianza modificada

$$\hat{s}^2 = \frac{N}{N-1}s^2$$

se encuentra que $\mu_{\hat{s}^2} = \sigma^2$, de manera que $\hat{s}^2$ es una estimación insesgada de σ^2. Sin embargo, $\hat{s}$ es una estimación sesgada de σ.

En el lenguaje de la esperanza matemática (ver capítulo 6) se puede decir que un estadístico es insesgado si su esperanza matemática es igual al correspondiente parámetro poblacional. Por lo tanto, $\bar{X}$ y $\hat{s}^2$ son insesgados, ya que $E\{\bar{X}\} = \mu$ y $E\{\hat{s}^2\} = \sigma^2$.

ESTIMACIONES EFICIENTES

Si la distribución muestral de dos estadísticos tiene la misma media (o esperanza), entonces al estadístico que tiene la menor varianza se le llama *estimador eficiente* del parámetro correspondiente, y al otro se le llama *estimador ineficiente*. A los valores de estos estadísticos se les llama *estimaciones eficientes e ineficientes*, respectivamente.

Si se consideran todos los estadísticos cuya distribución muestral tiene una misma media, al estadístico que tiene la menor varianza suele llamársele *estimador más eficiente* o *mejor* del parámetro correspondiente.

EJEMPLO 3 Las distribuciones muestrales de la media y de la mediana tienen la misma media, a saber, la media poblacional. Sin embargo, la varianza de la distribución muestral de las medias es menor que la varianza de la distribución muestral de las medianas (ver tabla 8.1). Por lo tanto, la media muestral proporciona una estimación eficiente de la media poblacional, en tanto que la mediana muestral proporciona una estimación ineficiente de la media poblacional.

De todos los estadísticos que estiman la media poblacional, la media muestral proporciona la mejor (o la más eficiente) estimación.

En la práctica, las estimaciones ineficientes suelen usarse debido a la relativa facilidad con que algunas de ellas pueden obtenerse.

ESTIMACIONES PUNTUALES Y ESTIMACIONES POR INTERVALO; SU CONFIABILIDAD

A una estimación de un parámetro poblacional que se da mediante un solo número se le llama *estimación puntual* del parámetro. A una estimación de un parámetro poblacional que se da mediante dos números, entre los cuales se considera que debe estar el parámetro en cuestión, se le llama *estimación por intervalo* del parámetro en cuestión.

Las estimaciones por intervalo dan la precisión, o exactitud, de la estimación, y por esto se prefieren a las estimaciones puntuales.

EJEMPLO 4 Si se dice que en la medición de una distancia se obtuvo como resultado 5.28 metros (m), se está dando una estimación puntual. En cambio, si se dice que la distancia es 5.28 ± 0.03 m (es decir, que la distancia está entre 5.25 y 5.31 m), se está dando una estimación por intervalo.

La información sobre el error (o precisión) de una estimación es su *confiabilidad*.

ESTIMACIÓN DE PARÁMETROS POBLACIONALES MEDIANTE UN INTERVALO DE CONFIANZA

Sean μ_S y σ_S la media y la desviación estándar (error estándar), respectivamente, de la distribución muestral de un estadístico S. Entonces, si la distribución muestral de S es aproximadamente normal (lo que se sabe que es así para muchos estadísticos si el tamaño de la muestra es $N \geq 30$), se puede esperar que exista un estadístico muestral S que se encuentre en los intervalos $\mu_S - \sigma_S$ a $\mu_S + \sigma_S$, $\mu_S - 2\sigma_S$ a $\mu_S + 2\sigma_S$ o $\mu_S - 3\sigma_S$ a $\mu_S + 3\sigma_S$, a 68.27%, 95.45% y 99.73% de las veces, respectivamente.

De igual manera, se puede hallar (o se puede tener *confianza* de hallar) μ_S en los intervalos $S - \sigma_S$ a $S + \sigma_S$, $S - 2\sigma_S$ a $S + 2\sigma_S$ o $S - 3\sigma_S$ a $S + 3\sigma_S$ a 68.27, 95.45 y 99.73% de las veces, respectivamente. Debido a ello, a estos intervalos se les llama *intervalos de confianza* de 68.27%, 95.45% y 99.73% para estimar μ_S. A los números de los extremos de estos intervalos ($S \pm \sigma_S$, $S \pm 2\sigma_S$ y $S \pm 3\sigma_S$) se les llama *límites de confianza* o *límites fiduciales*.

De igual manera, $S \pm 1.96\sigma_S$ y $S \pm 2.58\sigma_S$ son los límites de confianza de 95% y de 99% (o de 0.95 y 0.99) para S. Al porcentaje de confianza se le suele llamar *nivel de confianza*. A los números 1.96, 2.58, etc., que aparecen en los límites de confianza, se les llama *coeficientes de confianza* o *valores críticos* y se denotan z_c. A partir de los niveles de confianza se pueden encontrar los coeficientes de confianza y viceversa.

En la tabla 9.1 se presentan los valores de z_c que corresponden a varios niveles de confianza que se usan en la práctica. Los valores de z_c para niveles de confianza que no estén en esta tabla se pueden encontrar en las tablas de áreas de la curva normal (ver apéndice II).

Tabla 9.1

Nivel de confianza	99.73%	99%	98%	96%	95.45%	95%	90%	80%	68.27%	50%
z_c	3.00	2.58	2.33	2.05	2.00	1.96	1.645	1.28	1.00	0.6745

Intervalos de confianza para las medias

Si el estadístico S es la media muestral $\bar{X}$, entonces los límites de confianza de 95 y 99% para la estimación de la media poblacional μ están dados por $\bar{X} \pm 1.96\sigma_{\bar{X}}$ y $\bar{X} \pm 2.58\sigma_{\bar{X}}$, respectivamente. En general, los límites de confianza están dados por $\bar{X} \pm z_c\sigma_{\bar{X}}$, donde z_c (que depende del nivel de confianza deseado) puede leerse en la tabla 9.1. Empleando los valores para $\sigma_{\bar{X}}$ obtenidos en el capítulo 8, se ve que los límites de confianza para la media poblacional están dados por

$$\bar{X} \pm z_c \frac{\sigma}{\sqrt{N}} \tag{1}$$

si el muestreo se hace ya sea de una población infinita o de una población finita, pero con reposición, y están dados por

$$\bar{X} \pm z_c \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} \tag{2}$$

si el muestreo se hace sin reposición de una población de tamaño finito N_p.

Por lo general no se conoce la desviación estándar poblacional σ; de manera que para obtener los límites de confianza anteriores, se usa la estimación muestral $\hat{s}$ o s. El resultado es satisfactorio si $N \geq 30$. Si $N < 30$, la aproximación es pobre y se debe emplear la teoría del muestreo para muestras pequeñas (ver capítulo 11).

Intervalos de confianza para proporciones

Si el estadístico S es la proporción de "éxitos" en una muestra de tamaño N obtenida de una población binomial en la que p es la proporción de éxitos (es decir, la probabilidad de éxito), entonces los límites de confianza para p están dados por $P \pm z_c\sigma_p$, donde P es la proporción de éxitos en una muestra de tamaño N. Empleando los valores para σ_p indicados en el capítulo 8 se ve que los límites de confianza para la proporción poblacional están dados por

$$P \pm z_c \sqrt{\frac{pq}{N}} = P \pm z_c \sqrt{\frac{p(1-p)}{N}} \tag{3}$$

si el muestreo se hace de una población infinita o de una población finita, pero con reposición, y están dados por

$$P \pm z_c \sqrt{\frac{pq}{N}} \sqrt{\frac{N_p - N}{N_p - 1}} \tag{4}$$

si el muestreo se hace sin reposición y de una población finita de tamaño N_p.

Para calcular estos límites de confianza se emplea la estimación muestral P para p, la que por lo general resulta satisfactoria siempre que $N \geq 30$. En el problema 9.12 se da un método más exacto para obtener estos límites de confianza.

Intervalos de confianza para diferencias y sumas

Si S_1 y S_2 son dos estadísticos muestrales con distribuciones aproximadamente normales, los límites de confianza para la diferencia entre los parámetros poblacionales correspondientes a S_1 y S_2 están dados por

$$S_1 - S_2 \pm z_c\sigma_{S_1-S_2} = S_1 - S_2 \pm z_c\sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2} \tag{5}$$

y los límites de confianza para la suma de los parámetros poblacionales están dados por

$$S_1 + S_2 \pm z_c\sigma_{S_1+S_2} = S_1 + S_2 \pm z_c\sqrt{\sigma_{S_1}^2 + \sigma_{S_2}^2} \tag{6}$$

siempre que las muestras sean independientes (ver capítulo 8).

Por ejemplo, los límites de confianza para la diferencia entre dos medias poblacionales, en el caso en que las poblaciones sean infinitas, están dados por

$$\bar{X}_1 - \bar{X}_2 \pm z_c\sigma_{\bar{X}_1-\bar{X}_2} = \bar{X}_1 - \bar{X}_2 \pm z_c\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} \tag{7}$$

donde $\bar{X}_1$, σ_1, N_1 y $\bar{X}_2$, σ_2, N_2 son las correspondientes medias, desviaciones estándar y tamaños de las dos muestras obtenidas de las poblaciones.

De igual manera, los límites de confianza para la diferencia entre dos proporciones poblacionales, si las poblaciones son infinitas, están dados por

$$P_1 - P_2 \pm z_c\sigma_{P_1-P_2} = P_1 - P_2 \pm z_c\sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}} \tag{8}$$

donde P_1 y P_2 son las dos proporciones muestrales, N_1 y N_2 son los tamaños de las dos muestras obtenidas de las poblaciones y p_1 y p_2 son las proporciones en las dos poblaciones (estimadas por P_1 y P_2).

Intervalos de confianza para desviaciones estándar

Los límites de confianza para la desviación estándar σ de una población distribuida normalmente, estimada a partir de una muestra con desviación estándar s, están dados por

$$s \pm z_c\sigma_s = s \pm z_c\frac{\sigma}{\sqrt{2N}} \tag{9}$$

empleando la tabla 8.1. Para calcular estos límites de confianza, se usa s o $\hat{s}$ para estimar σ.

ERROR PROBABLE

Los límites de confianza de 50% para el parámetro poblacional correspondiente a un estadístico S están dados por $S \pm 0.6745\sigma_S$. La cantidad $0.6745\sigma_S$ se conoce como el *error probable* de la estimación.

PROBLEMAS RESUELTOS

ESTIMADORES INSESGADOS Y EFICIENTES

9.1 Dar un ejemplo de estimadores (o estimaciones) que sean: *a*) insesgados y eficientes, *b*) insesgados e ineficientes y *c*) sesgados e ineficientes.

SOLUCIÓN

a) La media muestral $\bar{X}$ y la varianza muestral

$$\hat{s}^2 = \frac{N}{N-1} s^2$$

son dos ejemplos.

b) La mediana muestral y el estadístico muestral $\frac{1}{2}(Q_1 + Q_3)$, donde Q_1 y Q_3 son los cuartiles muestrales inferior y superior, son dos de estos ejemplos. Ambos estadísticos son estimaciones insesgadas de la media poblacional, ya que la media de sus distribuciones muestrales es la media poblacional.

c) La desviación estándar s, la desviación estándar modificada $\hat{s}$, la desviación media y el rango semiintercuartil son cuatro de estos ejemplos.

9.2 Para el diámetro de un esfera, un científico obtiene una muestra de cinco mediciones, 6.33, 6.37, 6.36, 6.32 y 6.37 centímetros (cm). Obténganse estimaciones insesgadas y eficientes de: *a*) la verdadera media y *b*) la verdadera varianza.

SOLUCIÓN

a) La estimación insesgada y eficiente de la verdadera media (es decir, de la media poblacional) es

$$\bar{X} = \frac{\sum X}{N} = \frac{6.33 + 6.37 + 6.36 + 6.32 + 6.37}{5} = 6.35\,\text{cm}$$

b) La estimación insesgada y eficiente de la verdadera varianza (es decir de la varianza poblacional) es

$$\begin{aligned}\hat{s}^2 &= \frac{N}{N-1} s^2 = \frac{\sum (X - \hat{X})^2}{N-1} \\ &= \frac{(6.33 - 6.35)^2 + (6.37 - 6.35)^2 + (6.36 - 6.35)^2 + (6.32 - 6.35)^2 + (6.37 - 6.35)^2}{5-1} \\ &= 0.00055\,\text{cm}^2\end{aligned}$$

Obsérvese que aunque $\hat{s} = \sqrt{0.00055} = 0.023$ cm es una estimación de la verdadera desviación estándar, esta estimación no es ni insesgada ni eficiente.

9.3 Supóngase que las estaturas de 100 estudiantes varones de la universidad XYZ representan una muestra aleatoria de las estaturas de los 1 546 estudiantes de esa universidad. Determinar estimaciones insesgadas y eficientes: *a*) para la verdadera media y *b*) para la verdadera varianza.

SOLUCIÓN

a) De acuerdo con el problema 3.22, la estimación insesgada y eficiente de la verdadera estatura media es $\bar{X} = 67.45$ pulgadas (in).

b) De acuerdo con el problema 4.17, la estimación insesgada y eficiente de la verdadera varianza es

$$\hat{s}^2 = \frac{N}{N-1} s^2 = \frac{100}{99}(8.5275) = 8.6136$$

Por lo tanto, $\hat{s} = \sqrt{8.6136} = 2.93$ in. Obsérvese que como N es grande, en esencia no hay diferencia entre s^2 y $\hat{s}^2$ o entre s y $\hat{s}$.

Obsérvese que no se empleó la corrección de Sheppard por agrupamiento. Si se emplea, se usa $s = 2.79$ in (ver problema 4.21).

9.4 Dar una estimación insesgada e ineficiente del verdadero diámetro medio de la esfera del problema 9.2.

SOLUCIÓN

La mediana es un ejemplo de estimación insesgada e ineficiente de la media poblacional. Para las cinco mediciones coordenadas de acuerdo con su magnitud, la mediana es 6.36 cm.

INTERVALOS DE CONFIANZA PARA MEDIAS

9.5 Encontrar los intervalos de confianza: *a*) de 95% y *b*) 99% para estimar la estatura media de los estudiantes de la universidad XYZ del problema 9.3.

SOLUCIÓN

a) Los límites de confianza del 95% son $\bar{X} \pm 1.96\sigma/\sqrt{N}$. Empleando $\bar{X} = 67.45$ in, y $\hat{s} = 2.93$ in como estimación de σ (ver problema 9.3), los límites de confianza son $67.45 \pm 1.96(2.93/\sqrt{100})$ o 67.45 ± 0.57 in. Por lo tanto, el intervalo de confianza del 95% para la media poblacional μ es 66.88 a 68.02 in, lo que se denota así $66.88 < \mu < 68.02$.

De manera que se puede decir que la probabilidad de que la media poblacional de las estaturas se encuentre entre 66.88 y 68.02 es aproximadamente de 95% o 0.95. Empleando símbolos se escribe $\Pr\{66.88 < \mu < 68.02\} = 0.95$. Esto equivale a decir que se tiene 95% *de confianza* en que la media poblacional (o verdadera media) se encuentre entre 66.88 y 68.02 in.

b) Los límites de confianza del 99% son $\bar{X} \pm 2.58\sigma/\sqrt{N} = \bar{X} \pm 2.58\hat{s}/\sqrt{N} = 67.45 \pm 2.58(2.93/\sqrt{100}) = 67.45 \pm 0.76$ in. Por lo tanto, el intervalo de confianza del 99% para la media poblacional μ es 66.69 a 68.21 in, lo que se denota así $66.69 < \mu < 68.21$.

Al obtener los intervalos de confianza anteriores se supuso que la población era infinita o tan grande que se podía considerar que las condiciones eran las mismas que en un muestreo con reposición. En el caso de poblaciones finitas, si el muestreo se hace sin reposición, se debe usar

$$\frac{\sigma}{\sqrt{N}}\sqrt{\frac{N_p - N}{N_p - 1}} \qquad \text{en lugar de} \qquad \frac{\sigma}{\sqrt{N}}$$

Sin embargo, se puede considerar que el factor

$$\sqrt{\frac{N_p - N}{N_p - 1}} = \sqrt{\frac{1\,546 - 100}{1\,546 - 1}} = 0.967$$

es prácticamente 1.0, por lo que no necesita usarse. Si se usa, los límites de confianza anteriores se convierten en 67.45 ± 0.56 in y 67.45 ± 0.73 in, respectivamente.

9.6 Una empresa tiene 5 000 árboles de navidad maduros y listos para ser cortados y vendidos. En forma aleatoria se seleccionan 100 de estos árboles y se miden sus alturas. En la tabla 9.2 se dan estas alturas en pulgadas. Emplear MINITAB para dar un intervalo de confianza de 95% para la altura media de los 5 000 árboles. Si estos árboles se venden a $2.40 por pie, dar un límite inferior y un límite superior para el valor de los 5 000 árboles.

SOLUCIÓN

El intervalo de confianza de MINITAB, que se da a continuación, indica que la altura media de los 5 000 árboles puede ir desde 57.24 a 61.20 pulgadas. El número total de pulgadas en los 5 000 árboles está entre $(57.24)(5\,000) = 286\,200$ y

Tabla 9.2

56	61	52	62	63	34	47	35	44	59
70	61	65	51	65	72	55	71	57	75
53	48	55	67	60	60	73	74	43	74
71	53	78	59	56	62	48	65	68	51
73	62	80	53	64	44	67	45	58	48
50	57	72	55	56	62	72	57	49	62
46	61	52	46	72	56	46	48	57	52
54	73	71	70	66	67	58	71	75	50
44	59	56	54	63	43	68	69	55	63
48	49	70	60	67	47	49	69	66	73

(61.20)(5 000) = 306 000. Si estos árboles se venden a $2.40 por pie, entonces el precio por pulgada es $0.2. El valor de los árboles está entre (286 000)(0.2) = $57 200 y (306 000)(0.2) = $61 200 con 95% de confianza (o de seguridad).

Despliege de datos

```
altura
     56   70   53   71   73   50   46   54   44
     48   61   61   48   53   62   57   61   73
     59   49   52   65   55   78   80   72   52
     71   56   70   62   51   67   59   53   55
     46   70   54   60   63   65   60   56   64
     56   72   66   63   67   34   72   60   62
     44   62   56   67   43   47   47   55   73
     48   67   72   46   58   68   49   35   71
     74   65   45   57   48   71   69   69   44
     57   43   68   58   49   57   75   55   66
     59   75   74   51   48   62   52   50   63
     73

MTB > cl desviación estándar
```

Columna desviación estándar

```
Desviación estándar de altura = 10.111
MTB > zintervalo 95% de confianza ds = 10.111 datos en cl
```

Intervalos de confianza

```
Sigma supuesta = 10.1

Variable       N     Media     DesvEst     SE media      95.0% CI
Altura       100     59.22       10.11          1.01     (57.24, 61.20)
```

9.7 En una encuesta a sacerdotes católicos, cada sacerdote informó de la cantidad de bautizos, bodas y funerales celebrados el año anterior. En la tabla 9.3 se presentan las respuestas obtenidas. Utilizar estos datos para construir un intervalo de confianza de 95% para μ, la media del número, por sacerdote, de bautizos, bodas y fune-

Tabla 9.3

32	44	48	35	34	29	31	61	37	41
31	40	44	43	41	40	41	31	42	45
29	40	42	51	16	24	40	52	62	41
32	41	45	24	41	30	42	47	30	46
38	42	26	34	45	58	57	35	62	46

rales celebrados el año anterior. Obtener el intervalo empleando la fórmula para intervalos de confianza y usar también el comando `Zinterval` de MINITAB para hallar este intervalo.

SOLUCIÓN

Una vez ingresados los datos de la tabla 9.3 en la columna 1 de la hoja de cálculo de MINITAB y de haberle dado "número" como nombre a esta columna, se dan los comandos para la media y la desviación estándar.

```
MTB > cl media
```

Columna media

```
Media de número = 40.261
MTB > cl desviación estándar
```

Columna desviación estándar

```
Desviación estándar de número = 9.9895
```

El error estándar de la media es igual a $9.9895/\sqrt{50} = 1.413$, el valor crítico es 1.96 y el margen de error de 95% es 1.96(1.413) = 2.769. El intervalo de confianza va de 40.261 − 2.769 = 37.492 a 40.261 + 2.769 = 43.030.

Con el comando `Zinterval` se obtiene el resultado siguiente:

```
MTB > Zinterval, de 95% de confianza sd = 9.9895 datos en cl
```

Intervalos de confianza Z

```
Sigma supuesta = 9.99

Variable    N    Media    DesvEst    SE media     95.00% CI
Número     50    40.26       9.99         1.41    (37.49, 43.03)
```

Se tiene una confianza de 95% de que la verdadera media de todos los sacerdotes esté entre 37.49 y 43.03.

9.8 Para medir el tiempo de reacción, un psicólogo estima que la desviación estándar es 0.05 segundos (s). ¿Qué tan grande debe ser la muestra de las medidas para que se tenga una confianza: *a*) de 95% y *b*) de 99% en que el error de esta estimación no será mayor de 0.01 s?

SOLUCIÓN

a) Los límites de confianza del 95% son $\bar{X} \pm 1.96\sigma/\sqrt{N}$, siendo el error de estimación $1.96\sigma/\sqrt{N}$. Tomando $\sigma = s =$ 0.05 s, se ve que este error será igual a 0.01 s si $(1.96)(0.05)/\sqrt{N} = 0.01$; es decir, $\sqrt{N} = (1.96)(0.05)/0.01 = 9.8$ o bien $N = 96.04$. Por lo tanto, se puede tener una confianza del 95% en que el error de estimación será menor a 0.01 si N es 97 o mayor.

Otro método

$$\frac{(1.96)(0.05)}{\sqrt{N}} \leq 0.01 \quad \text{si} \quad \frac{\sqrt{N}}{(1.96)(0.05)} \geq \frac{1}{0.01} \quad \text{o bien} \quad \sqrt{N} \geq \frac{(1.96)(0.05)}{0.01} = 9.8$$

Entonces $N \geq 96.04$, o bien $N \geq 97$.

b) Los límites de confianza del 99% son $\bar{X} \pm 2.58\sigma/\sqrt{N}$. Entonces $(2.58)(0.05)/\sqrt{N} = 0.01$ o bien $N = 166.4$. De manera que se puede tener una confianza de 99% de que el error de estimación será menor a 0.01 s sólo si N es 167 o mayor.

9.9 De un total de 200 calificaciones de matemáticas se tomó una muestra aleatoria de 50 calificaciones en la que la media encontrada fue 75 y la desviación estándar, 10.

a) ¿Cuáles son los límites de confianza de 95% para la estimación de la media de las 200 calificaciones?

b) ¿Con qué grado de confianza se puede decir que la media de las 200 calificaciones es 75 ± 1?

SOLUCIÓN

a) Como el tamaño de la población no es muy grande en comparación con el tamaño de la muestra, hay que hacer un ajuste. Entonces, los límites de confianza de 95% son

$$\bar{X} \pm 1.96\sigma_{\bar{X}} = \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} = 75 \pm 1.96 \frac{10}{\sqrt{50}} \sqrt{\frac{200 - 50}{200 - 1}} = 75 \pm 2.4$$

b) Los límites de confianza están representados por

$$\bar{X} \pm z_c \sigma_{\bar{X}} = \bar{X} \pm z_c \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}} = 75 \pm z_c \frac{10}{\sqrt{50}} \sqrt{\frac{200 - 50}{200 - 1}} = 75 \pm 1.23 z_c$$

Como esto debe ser igual a 75 ± 1, se tiene $1.23z_c = 1$, o bien $z_c = 0.81$. El área bajo la curva normal desde $z = 0$ hasta $z = 0.81$ es 0.2910; por lo tanto, el grado de confianza buscado es $2(0.2910) = 0.582$ o bien 58.2%.

INTERVALOS DE CONFIANZA PARA PROPORCIONES

9.10 Un sondeo realizado con 100 votantes tomados en forma aleatoria de la población de todos los votantes de determinado distrito indica que de éstos, 55% están a favor de cierto candidato. Encontrar límites de confianza de: *a*) 95%, *b*) 99% y *c*) 99.73% para la proporción de todos los votantes a favor de este candidato.

SOLUCIÓN

a) Los límites de confianza de 95% para la p poblacional son $P \pm 1.96\sigma_P = P \pm 1.96\sqrt{p(1-p)/N} = 0.55 \pm 1.96\sqrt{(0.55)(0.45)/100} = 0.55 \pm 0.10$, donde se ha usado la proporción muestral P para estimar p.

b) Los límites de confianza de 99% para p son $0.55 \pm 2.58\sqrt{(0.55)(0.45)/100} = 0.55 \pm 0.13$.

c) Los límites de confianza de 99.73% para p son $0.55 \pm 3\sqrt{(0.55)(0.45)/100} = 0.55 \pm 0.15$.

9.11 ¿De qué tamaño deberá tomarse la muestra de votantes del problema 9.19 para tener una confianza de: *a*) 95% y *b*) 99.73% de que el candidato será electo?

SOLUCIÓN

Los límites de confianza para p son $P \pm z_c\sqrt{p(1-p)/N} = 0.55 \pm z_c\sqrt{(0.55)(0.45)/N} = 0.55 \pm 0.50z_c/\sqrt{N}$, donde, de acuerdo con el problema 9.10, se ha usado la estimación $P = p = 0.55$. Dado que el candidato gana sólo si tiene más del 50% de la población de votantes, se requiere que $0.50z_c/\sqrt{N}$ sea menor a 0.05.

a) Para una confianza de 95%, $0.50z_c/\sqrt{N} = 0.50(1.96)/\sqrt{N} = 0.05$ si $N = 384.2$. Por lo tanto, N debe ser 385, por lo menos.

b) Para una confianza de 99.73%, $0.50z_c/\sqrt{N} = 0.50(3)/\sqrt{N} = 0.05$ si $N = 900$. Por lo tanto, N debe ser 901, por lo menos.

Otro método

$1.50/\sqrt{N} < 0.05$ si $\sqrt{N}/1.50 > 1/0.05$ o $\sqrt{N} > 1.50/0.05$. Entonces $\sqrt{N} > 30$ o bien $N > 900$, de manera que N debe ser por lo menos 901.

9.12 Se realiza un estudio y se encuentra que 156 de 500 varones adultos son fumadores. Emplear el paquete de software STATISTIX para dar un intervalo de confianza de 99% para p, la proporción poblacional de varones adultos que son fumadores. Verificar el intervalo de confianza calculándolo a mano.

SOLUCIÓN

Los resultados de STATISTIX se dan a continuación. El intervalo de confianza de 99% aparece en negritas.

Prueba de proporción de una muestra

```
Tamaño de la muestra         500
Éxito                        156
Proporción               0.31200

Hipótesis nula           P = 0.5
Hipótesis alterna      P < > 0.5

Diferencia              -0.18800
Error estándar           0.02072
Z (sin corregir)           -8.41     P  0.0000
Z (corregida)              -8.36     P  0.0000

                 Intervalo de confianza 99%
Sin corregir         (0.25863, 0.36537)
Corregido            (0.25763, 0.36637)
```

Se tiene una confianza de 99% de que el verdadero porcentaje de varones adultos fumadores esté entre 25.9% y 36.5%.

Verificación:

$$P = 0.312,\ z_c = 2.58,\ \sqrt{\frac{0.312(0.688)}{500}} = 0.0207$$

$P \pm z_c\sqrt{\dfrac{p(1-p)}{N}}$ o bien $0.312 \pm 2.58(0.0207)$ o bien (**0.258**, **0.365**). Esto es lo mismo que se obtuvo antes con el paquete de software STATISTIX.

9.13 Refiérase al problema 9.12 para dar un intervalo de confianza de 99% para p empleando MINITAB.

SOLUCIÓN

El intervalo de confianza de 99% se muestra abajo en negritas. Es el mismo que el intervalo de confianza obtenido con STATISTIX en el problema 9.12.

Muestra	X	N	Muestra P	CI 99%	Valor z	Valor P
1	156	500	0.312000	**(0.258629, 0.365371)**	-8.41	0.000

INTERVALOS DE CONFIANZA PARA DIFERENCIAS Y SUMAS

9.14 Para comparar la cantidad de tiempo que utilizan su celular los estudiantes universitarios, tanto varones como mujeres, se tomaron 50 estudiantes varones y 50 estudiantes mujeres y se determinó la cantidad de tiempo, en horas por semana, que utilizan su celular. En la tabla 9.4 se presentan los resultados en horas. Dar un intervalo de 95% de confianza para $\mu_1 - \mu_2$ usando MINITAB. Verificar los resultados calculando a mano el intervalo.

Tabla 9.4

Varones					Mujeres				
12	4	11	13	11	11	9	7	10	9
7	9	10	10	7	10	10	7	9	10
7	12	6	9	15	11	8	9	6	11
10	11	12	7	8	10	7	9	12	14
8	9	11	10	9	11	12	12	8	12
10	9	9	7	9	12	9	10	11	7
11	7	10	10	11	12	7	9	8	11
9	12	12	8	13	10	8	13	8	10
9	10	8	11	10	9	9	9	11	9
13	13	9	10	13	9	8	9	12	11

SOLUCIÓN

Dado que ambas muestras son mayores de 30, se puede usar indistintamente la prueba *z* o la prueba *t* para dos muestras, ya que la distribución *t* y la distribución *z* son muy similares.

```
Dos muestras T para varones vs mujeres
            N     Media     DesvEst    SE media
varones    50      9.82       2.15         0.30
mujeres    50      9.70       1.78         0.25
Diferencia = mu (varones) - mu (mujeres)
Estimado para diferencia: 0.120000
CI 95% para diferencia: (-0.663474, 0.903474)
Prueba T de diferencia = 0 (vs no =): valor T = 0.30 valor P = 0.762
DF = 98
Ambos utilizaron la desviación estándar común = 1.9740
```

De acuerdo con los resultados de MINITAB, la diferencia entre las medias poblacionales está entre -0.66 y 0.90. Así que existe la posibilidad de que no haya diferencia entre estas medias poblacionales.

Verificación:

La fórmula para un intervalo de confianza de 95% es $(\bar{x}_1 - \bar{x}_2) \pm z_c\left(\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}\right)$. Sustituyendo se obtiene $0.12 \pm 1.96(0.395)$ que corresponde a la respuesta dada por MINITAB.

9.15 Usar STATISTIX y SPSS para resolver el problema 9.14.

SOLUCIÓN

A continuación se presenta la solución dada por STATISTIX. Obsérvese que el intervalo de confianza de 95% es el mismo que el del problema 9.14. Más adelante se verá por qué se supone que las varianzas son iguales.

```
Pruebas de dos muestras T para varones vs mujeres

Variable    Media     N       SD        SE
varones    9.8200    50   2.1542    0.3046
mujeres    9.7000    50   1.7757    0.2511
Diferencia 0.1200

Hipótesis nula: diferencia = 0
Hipótesis alterna: diferencia < > 0

                                                CI 95% para diferencia
Supuesto                  T       DF        P   Inferior     Superior
Varianzas iguales      0.30       98   0.7618    -0.6635       0.9035
Varianzas desiguales   0.30     94.6   0.7618    -0.6638       0.9038

Prueba para igualdad          F          DF          P
        de varianzas       1.47       49,49      0.0899
```

La solución dada por SPSS es la siguiente:

Grupo estadístico

	Sexo	N	Media	Desviación estándar	Media de error estándar
momento	1.00	50	9.7000	1.77569	.25112
	2.00	50	9.8200	2.15416	.30464

Prueba de muestras independientes

	Prueba de Levene para igualdad de varianzas		Prueba t para igualdad de medias						
								Intervalo de confianza de 95% de la diferencia	
	F	Sig.	t	gl	Sig. (2-finales)	Diferencia media	Diferencia error estándar	Inferior	Superior
momento Varianzas iguales supuestas	.898	.346	−.304	98	.762	−.12000	.39480	−.90347	.66347
Varianzas iguales no supuestas			−.304	94.556	.762	−.12000	.39480	−.90383	.66383

9.16 Usar SAS para resolver el problema 9.14. Dar las formas de archivos de datos que permiten usar SAS para realizar este análisis.

SOLUCIÓN

El análisis de SAS es como se muestra a continuación. El intervalo de confianza se ha impreso en negritas en la parte inferior de los resultados.

```
Dos muestras: prueba t para las medias de varones y mujeres

Estadísticos de muestra

Grupo      N     Media    DesvEst     ErrorEst
-----------------------------------------------
varones   50      9.82     2.1542       0.3046
mujeres   50      9.7      1.7757       0.2511

Hipótesis nula: media 1 - media 2 = 0
Alternativa     Media 1 - media 2 ^= 0

Si las varianzas son      estadístico t    Df     Pr > t
-----------------------------------------------------------
Igual                     0.304         98        0.7618
Desigual                  0.304         94.56     0.7618

Intervalo de confianza 95% para la diferencia entre dos medias.

            Límite inferior     Límite superior
            ---------------     ---------------
                 -0.66               0.90
```

Los archivos de datos que se emplean con SAS para el análisis pueden tener los datos de varones y de mujeres en columnas separadas, pero los datos también pueden consistir en las horas que se emplea el celular, en una columna, y el sexo de la persona (varón o mujer), en otra columna. Varones y mujeres se pueden codificar como 1 y 2, respectivamente. En la primera forma habrá 2 columnas y 50 renglones. En la segunda forma habrá 2 columnas y 100 renglones.

INTERVALOS DE CONFIANZA PARA DESVIACIONES ESTÁNDAR

9.17 Para un intervalo de confianza para la varianza de una población se utiliza la distribución Ji cuadrada. El intervalo de confianza $(1 - \alpha) \times 100\%$ es $\frac{(n-1)S^2}{(\chi^2_{\alpha/2})} < \sigma^2 < \frac{(n-1)S^2}{(\chi^2_{1-\alpha/2})}$ donde n es el tamaño de la muestra, S^2 es la varianza muestral, $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ pertenecen a la distribución Ji cuadrada con $(n - 1)$ grados de libertad. Use EXCEL para hallar un intervalo de confianza de 99% para la varianza de veinte recipientes de 180 onzas. Los datos de los veinte recipientes se presentan en la tabla 9.5.

Tabla 9.5

181.5	180.8
179.7	182.4
178.7	178.5
183.9	182.2
179.7	180.9
180.6	181.4
180.4	181.4
178.5	180.6
178.8	180.1
181.3	182.2

SOLUCIÓN

A continuación se presenta la hoja de cálculo de EXCEL. Los datos se encuentran en A1:B10. En la columna D se muestran las funciones cuyos valores aparecen en la columna C.

A	B	C	D
181.5	180.8	2.154211	=VAR(A1:B10)
179.1	182.4	40.93	=19*C1
178.7	178.5	38.58226	=CHIINV(0.005,19)
183.9	182.2	6.843971	=CHIINV(0.995,19)
179.7	180.9		
180.6	181.4	1.06085	=C2/C3
180.4	181.4	5.980446	=C2/C4
178.5	180.6		
178.8	180.1		
181.3	182.2		

El intervalo de confianza de 99% para σ^2 es: $(1.06085 < \sigma^2 < 5.980446)$. El intervalo de confianza de 99% para σ es: (1.03, 2.45).

Obsérvese que con =VAR(A1:B10) se obtiene S^2, con =CHIINV(0.005,19) se obtiene el valor de Ji cuadrada que tiene a su derecha un área de 0.005, y con =CHIINV(0.995,19) el valor de ji cuadrada que tiene a su derecha un área de 0.995. En ambos casos, la distribución ji cuadrada tiene 19 grados de libertad.

9.18 Para comparar la varianza de una población con la varianza de otra población se emplea el siguiente intervalo de confianza $(1 - \alpha) \times 100\%$:

$$\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{\alpha/2(\nu_1,\nu 2)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} F_{\alpha/2(\nu_2,\nu_1)},$$

donde n_1 y n_2 son los tamaños de las dos muestras, S_1^2 y S_2^2 son las dos varianzas muestrales, $v_1 = n_1 - 1$ y $v_2 = n_2 - 1$ son los grados de libertad, en el numerador y en el denominador, para la distribución F y los valores F pertenecen a la distribución F. En la tabla 9.6 se dan los números de correos electrónicos enviados por semana por los empleados de dos empresas.

Dar un intervalo de confianza de 95% para $\frac{\sigma_1}{\sigma_2}$.

Tabla 9.6

Empresa 1	Empresa 2
81	99
104	100
115	104
111	98
85	103
121	113
95	95
112	107
100	98
117	95
113	101
109	109
101	99
	93
	105

SOLUCIÓN

A continuación se muestra la hoja de cálculo de EXCEL. En la columna D se muestran las funciones cuyos valores aparecen en la columna C. En C1 y C2 se calculan las dos varianzas muestrales. En C3 y C4 se calculan los valores F. En C5 y C6 se calculan los extremos del intervalo de confianza para el cociente de las varianzas. Como se ve, el intervalo de confianza del 95% para $\frac{\sigma_1^2}{\sigma_2^2}$ es (1.568, 15.334). El intervalo de confianza del 95% para $\frac{\sigma_1}{\sigma_2}$ es (1.252, 3.916). Obsérvese que =FINV(0.025,12,14) es el punto que corresponde a la distribución F, con $\nu_1 = 12$ y $\nu_2 = 14$ grados de libertad, que tiene un área de 0.025 a su derecha.

A	B	C	D
Compañía 1	Compañía 2	148.5769231	=VAR(A2:A14)
81	99	31.06666667	=VAR(B2:B16)
104	100	3.050154789	=FINV(0.025,12,14)
115	104	3.2062117	=FINV(0.025,14,12)
111	98	1.567959436	=(C1/C2)/C3
85	103	15.33376832	=(C1/C2)*C4
121	113		
95	95	1.25218187	=SQRT(C5)
112	107	3.915835584	=SQRT(C6)
100	98		
117	95		
113	101		
109	109		
101	99		
	93		
	105		

ERROR PROBABLE

9.19 La media del voltaje de 50 baterías del mismo tipo es 18.2 volts (V) y la desviación estándar es 0.5 V. Encontrar: *a*) el error probable de la media y *b*) los límites de confianza de 50%.

SOLUCIÓN

a)
$$\text{Error probable de la media} = 0.674\sigma_{\hat{X}} = 0.6745\frac{\sigma}{\sqrt{N}} = 0.6745\frac{\hat{s}}{\sqrt{N}}$$
$$= 0.6745\frac{s}{\sqrt{N-1}} = 0.6745\frac{0.5}{\sqrt{49}} = 0.048 \text{ V}$$

Obsérvese que si la desviación estándar de 0.5 V se calcula como $\hat{s}$, el error probable también es $0.6745(0.5/\sqrt{50}) = 0.048$, de manera que si N es suficientemente grande puede usarse cualquier estimación.

b) Los límites de confianza de 50% son 18 ± 0.048 V.

9.20 Una medición se registra como 216.480 gramos (g) con un error probable de 0.272 g. ¿Cuáles son los límites de confianza de 95% para esta medición?

SOLUCIÓN

El error probable es $0.272 = 0.6745\sigma_{\bar{X}}$ o bien $\sigma_{\bar{X}} = 0.272/0.6745$. Por lo tanto, los límites de confianza de 95% son $\bar{X} \pm 1.96\sigma_{\bar{X}} = 216.480 \pm 1.96(0.272/0.6745) = 216.480 \pm 0.790$ g.

PROBLEMAS SUPLEMENTARIOS

ESTIMADORES INSESGADOS Y EFICIENTES

9.21 Las mediciones de una muestra de masas fueron 8.3, 10.6, 9.7, 8.8, 10.2 y 9.4 kilogramos (kg), respectivamente. Determinar estimaciones insesgadas y eficientes de: *a*) la media poblacional, *b*) la varianza poblacional y *c*) comparar la desviación estándar muestral con la desviación estándar poblacional estimada.

9.22 En una muestra de 10 cinescopios de televisión, producidos por una empresa, la media del tiempo de vida es 1 200 horas (h) y la desviación estándar es 100 h. Estimar: *a*) la media y *b*) la desviación estándar de todos los cinescopios producidos por esta empresa.

9.23 *a*) Repetir el problema 9.22 considerando que la muestra es de 30, 50 y 100 cinescopios de televisión.

b) ¿Qué se puede concluir sobre la relación entre la desviación estándar muestral y las estimaciones de la desviación estándar poblacional obtenidas con diferentes tamaños de muestra?

INTERVALOS DE CONFIANZA PARA MEDIAS

9.24 La media y la desviación estándar de la carga máxima que soporta cada uno de 60 cables (ver problema 3.59) son 11.09 toneladas y 0.73 toneladas, respectivamente. Encontrar los límites de confianza: *a*) de 95% y *b*) de 99% para la media de la carga máxima de cada uno de los cables producidos por la empresa.

9.25 La media y la desviación estándar de los diámetros de una muestra de 250 cabezas de remaches fabricados por una empresa son 0.72643 in y 0.00058 in, respectivamente (ver problema 3.61). Encontrar los límites de confianza de: *a*) 99%, *b*) 98%, *c*) 95% y *d*) 90% para los diámetros de todas las cabezas de remaches producidos por la empresa.

9.26 Encontrar: *a*) los límites de confianza de 50% y *b*) el error probable para la media de los diámetros del problema 9.25.

9.27 Si se estima que la desviación estándar del tiempo de vida de los cinescopios de televisión es de 100 h, ¿de qué tamaño deberá tomarse la muestra para que se tenga una confianza de: *a*) 95%, *b*) 90%, *c*) 99% y *d*) 99.73% de que el error en la vida media estimada no sea mayor de 20 h?

9.28 A los integrantes de un grupo de 50 personas que acostumbra comprar por Internet se les preguntó cuánto gastaban anualmente en estas compras por Internet. Las respuestas obtenidas se presentan en la tabla 9.7.

Empleando las ecuaciones del capítulo 9, así como paquetes de software para estadística, encontrar un intervalo de 80% para μ, la cantidad media gastada por las personas que compran por Internet.

Tabla 9.7

418	379	77	212	378
363	434	348	245	341
331	356	423	330	247
351	151	220	383	257
307	297	448	391	210
158	310	331	348	124
523	356	210	364	406
331	364	352	299	221
466	150	282	221	432
366	195	96	219	202

9.29 Una empresa tiene 500 cables. En una prueba realizada a 40 cables tomados en forma aleatoria se encuentra que la resistencia media a la ruptura es 2 400 libras (lb) y la desviación estándar es 150 lb.

a) ¿Cuáles son los límites de confianza de 95% y 99% para la estimación de la resistencia media a la ruptura de los 460 cables restantes?

b) ¿Con qué grado de confianza se puede decir que la resistencia media a la ruptura de los 460 cables restantes es 2 400 $\pm$ 35 lb?

INTERVALOS DE CONFIANZA PARA PROPORCIONES

9.30 Una urna contiene canicas rojas y blancas en proporción desconocida. En una muestra aleatoria de 60 canicas tomadas de esta urna, con reposición, se observó que 70% eran rojas. Encontrar límites de confianza de: *a*) 95%, *b*) 99% y *c*) 99.73% para la verdadera proporción de canicas rojas en esta urna.

9.31 Se realizó un sondeo con 1 000 personas mayores de 65 años para determinar el porcentaje de la población de este grupo de edad que tiene conexión a Internet. Se encontró que 387 de las 1 000 personas contaban con conexión a Internet. Empleando las ecuaciones dadas en este libro, así como software para estadística, encontrar un intervalo de confianza de 97.5% para p.

9.32 Se cree que los resultados de la elección entre dos candidatos sean muy reñidos. ¿Cuál será la cantidad mínima de votantes que habrá que sondear para tener una confianza de: *a*) 80%, *b*) 90%, *c*) 95% y *d*) 99% para una decisión a favor de cualquiera de los candidatos?

INTERVALOS DE CONFIANZA PARA DIFERENCIAS Y SUMAS

9.33 Se tienen dos grupos similares de pacientes, *A* y *B*, que constan de 50 y 100 individuos, respectivamente. A las personas del primer grupo se les administra una nueva pastilla para dormir, y a las del segundo, una pastilla convencional. En los pacientes del grupo *A* la media de la cantidad de horas de sueño es 7.82 y la desviación estándar 0.24 h; en los pacientes del grupo *B* la media de la cantidad de horas de sueño es 6.75 y la desviación estándar es 0.30 h. Encontrar los límites de confianza: *a*) de 95% y *b*) de 99% para la diferencia entre las medias de la cantidad de horas de sueño inducido por los dos tipos de pastillas para dormir.

9.34 Se realiza un estudio para comparar la duración media de vida de los varones con la de las mujeres. Se toman muestras aleatorias de las páginas del obituario; los datos recolectados se presentan en la tabla 9.8.

Usando los resultados proporcionados en dicha tabla, las ecuaciones presentadas en este libro y un software para estadística, dar un intervalo de confianza de 85% para $\mu_{\text{VARONES}} - \mu_{\text{MUJERES}}$.

Tabla 9.8

Varones					Mujeres				
85	53	100	49	65	64	93	82	71	77
60	51	61	83	65	64	60	75	87	60
55	99	56	55	55	61	84	91	61	85
90	72	62	69	59	105	90	59	86	62
49	72	58	60	68	71	99	98	54	94
90	74	85	80	77	98	61	108	79	50
62	65	81	55	71	66	74	60	90	95
78	49	78	80	75	81	86	65	86	81
53	82	109	87	78	92	77	82	86	79
72	104	70	31	50	91	93	63	93	53

9.35 Se comparan dos áreas de un país respecto a la proporción de adolescentes con caries. En una de estas áreas se agrega flúor al agua y en la otra no. En la muestra del área en donde no se agrega flúor al agua, 425 de 1 000 adolescentes tienen por lo menos una caries. En la muestra del área en donde sí se agrega flúor al agua, 376 de 1 000 adolescentes tienen por lo menos una caries. Dar un intervalo de confianza de 99% para esta diferencia, en porcentaje, empleando las ecuaciones dadas en este libro, así como un paquete de software para estadística.

INTERVALOS DE CONFIANZA PARA DESVIACIONES ESTÁNDAR

9.36 La desviación estándar en la resistencia a la ruptura encontrada en 100 cables de una empresa es 180 lb. Dar límites de confianza de: *a*) 95%, *b*) 99% y *c*) 99.73% para la desviación estándar de todos los cables producidos por esta empresa.

9.37 Resolver el problema 9.17 empleando SAS.

9.38 Resolver el problema 9.18 empleando SAS.

TEORÍA ESTADÍSTICA DE LA DECISIÓN

10

DECISIONES ESTADÍSTICAS

En la práctica, con frecuencia se tienen que tomar decisiones acerca de una población con base en información muestral. A tales decisiones se les llama *decisiones estadísticas*. Por ejemplo, tal vez se tenga que decidir, con base en datos muestrales, si determinado suero es realmente eficaz en la curación de una enfermedad, si un método educativo es mejor que otro, o bien si una moneda está alterada o no.

HIPÓTESIS ESTADÍSTICAS

Cuando se trata de tomar una decisión es útil hacer suposiciones (o conjeturas) acerca de la población de que se trata. A estas suposiciones, que pueden ser o no ciertas, se les llama *hipótesis estadísticas*. Estas hipótesis estadísticas son por lo general afirmaciones acerca de las distribuciones de probabilidad de las poblaciones.

Hipótesis nula

En muchas ocasiones se formula una hipótesis estadística con la única finalidad de refutarla o anularla. Por ejemplo, si se quiere decidir si una moneda está cargada o no, se formula la hipótesis de que no está cargada (es decir, $p = 0.5$, donde p es la probabilidad de cara). También, si se quiere decidir si un método es mejor que otro, se formula la hipótesis de que *no hay diferencia* entre los dos (es decir, que cualquier diferencia que se observe se debe sólo a las fluctuaciones del muestreo de una misma población). A estas hipótesis se les llama *hipótesis nula* y se denota H_0.

Hipótesis alternativa

A toda hipótesis que difiera de la hipótesis dada se le llama *hipótesis alternativa*. Por ejemplo, si una hipótesis es $p = 0.5$, la hipótesis alternativa puede ser $p = 0.7$, $p \neq 0.5$ o $p > 0.5$. La hipótesis alternativa a la hipótesis nula se denota H_1.

PRUEBAS DE HIPÓTESIS Y DE SIGNIFICANCIA O REGLAS DE DECISIÓN

Si se supone que una hipótesis es verdadera, pero se encuentra que los resultados que se observan en una muestra aleatoria difieren marcadamente de los resultados esperados de acuerdo con la hipótesis (es decir, esperados con base sólo en la casualidad, empleando la teoría del muestreo), entonces se dice que las diferencias observadas son *significativas* y se estará inclinado a rechazar la hipótesis (o por lo menos a no aceptarla de acuerdo con la evidencia obtenida). Por ejemplo, si en 20 lanzamientos de una moneda se obtienen 16 caras, se estará inclinado a rechazar que la moneda es buena, aun cuando se puede estar equivocado.

A los procedimientos que permiten determinar si las muestras observadas difieren significativamente de los resultados esperados, ayudando así a decidir si se acepta o se rechaza la hipótesis, se les llama *pruebas de hipótesis*, *pruebas de significancia* o *reglas de decisión*.

ERRORES TIPO I Y TIPO II

Si se rechaza una hipótesis que debería aceptarse se dice que se comete un *error tipo I*. Si por otro lado, se acepta una hipótesis que debería rechazarse, se comete un *error tipo II*. En cualquiera de los casos ha habido una decisión errónea o se ha hecho un juicio erróneo.

Para que las reglas de decisión (o pruebas de hipótesis) sean buenas, deben diseñarse de manera que se minimicen los errores de decisión. Esto no es sencillo, ya que para cualquier tamaño dado de muestra, al tratar de disminuir un tipo de error suele incrementarse el otro tipo de error. En la práctica, un tipo de error puede ser más importante que otro y habrá que sacrificar uno con objeto de limitar al más notable. La única manera de reducir los dos tipos de error es aumentando el tamaño de la muestra, lo que no siempre es posible.

NIVEL DE SIGNIFICANCIA

Cuando se prueba determinada hipótesis, a la probabilidad máxima con la que se está dispuesto a cometer un error tipo I se le llama *nivel de significancia* de la prueba. Esta probabilidad acostumbra denotarse α y por lo general se especifica antes de tomar cualquier muestra para evitar que los resultados obtenidos influyan sobre la elección del valor de esta probabilidad.

En la práctica, se acostumbran los niveles de significancia 0.05 o 0.01, aunque también se usan otros valores. Si, por ejemplo, al diseñar la regla de decisión se elige el nivel de significancia 0.05 (o bien 5%), entonces existen 5 posibilidades en 100 de que se rechace una hipótesis que debía ser aceptada; es decir, se tiene una *confianza* de aproximadamente 95% de que se ha tomado la decisión correcta. En tal caso se dice que la hipótesis ha sido rechazada al nivel de significancia 0.05, lo que significa que la hipótesis tiene una probabilidad de 0.05 de ser errónea.

PRUEBAS EMPLEANDO DISTRIBUCIONES NORMALES

Para ilustrar las ideas presentadas antes, supóngase que de acuerdo con determinada hipótesis, la distribución muestral de un estadístico S es una distribución normal con media μ_S y desviación estándar σ_S. Por lo tanto, la distribución de la variable estandarizada (o puntuación z), dada por $z = (S - \mu_S)/\sigma_S$, es la distribución normal estándar (media 0, varianza 1), que se muestra en la figura 10-1.

Como indica la figura 10-1, se puede tener una confianza del 95% en que si la hipótesis es verdadera, entonces la puntuación z del estadístico muestral real S estará entre -1.96 y 1.96 (ya que el área bajo la curva normal entre estos dos valores es 0.95). Pero si se toma una sola muestra aleatoria y se encuentra que la puntuación z del estadístico se encuentra *fuera* del rango -1.96 a 1.96, se concluye que si la hipótesis dada es verdadera, esto sólo puede ocurrir con una probabilidad de 0.05 (el total del área sombreada en la figura). En tal caso se dice que la puntuación z difiere en forma *significativa* de lo esperado de acuerdo con la hipótesis dada y se estará inclinado a rechazar esa hipótesis.

El 0.05, que es el total de área sombreada, es el nivel de significancia de la prueba. Esta cantidad representa la probabilidad de estar equivocado al rechazar la hipótesis (es decir, la probabilidad de cometer un error tipo I). Por lo tanto, se dice que la hipótesis *se rechaza al nivel de significancia 0.05* o que la puntuación z del estadístico muestral dado es *significante al nivel 0.05*.

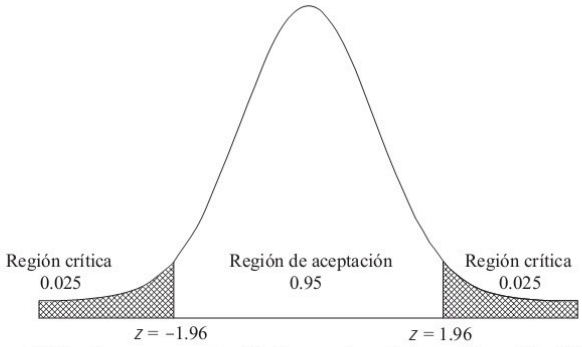


Figura 10-1 Curva normal estándar mostrando la región crítica (0.05) y la región de aceptación (0.95).

El conjunto de puntuaciones z que queda fuera del intervalo -1.96 a 1.96 constituye lo que se llama *región crítica de la hipótesis*, *región de rechazo de la hipótesis* o *región de significancia*. Al conjunto de puntuaciones z que queda dentro del intervalo -1.96 a 1.96 se le llama *región de aceptación de la hipótesis* o *región de no significancia*.

De acuerdo con las observaciones anteriores, se puede formular la siguiente regla de decisión (o prueba de hipótesis o de significancia):

Rechazar la hipótesis, al nivel de significancia 0.05, si la puntuación z del estadístico S se encuentra fuera del rango -1.96 a 1.96 (es decir, si $z > 1.96$ o $z < -1.96$). Esto equivale a decir que el estadístico muestral observado es significante al nivel 0.05.

Si no es así, se acepta la hipótesis (o, si se desea, no se toma ninguna decisión).

Debido a que la puntuación z es tan importante en las pruebas de hipótesis, también se le conoce como el *estadístico de prueba*.

Hay que hacer notar que también pueden emplearse otros niveles de significancia. Por ejemplo, si se emplea el nivel 0.01, el 1.96, empleado antes se sustituirá por 2.58 (ver la tabla 10.1). También se puede emplear la tabla 9.1, ya que los niveles de significancia y de confianza suman 100%.

Tabla 10.1

Nivel de significancia, α	0.10	0.05	0.01	0.005	0.002
Valores críticos de z para pruebas de una cola	-1.28 *o* 1.28	-1.645 *o* 1.645	-2.33 *o* 2.33	-2.58 *o* 2.58	-2.88 *o* 2.88
Valores críticos de z para pruebas de dos colas	-1.645 *y* 1.645	-1.96 *y* 1.96	-2.58 *y* 2.58	-2.81 *y* 2.81	-3.08 *y* 3.08

PRUEBAS DE UNA Y DE DOS COLAS

En la prueba anterior interesaban los valores extremos del estadístico S, o de sus correspondientes puntuaciones z, a *ambos* lados de la media (es decir, en las dos colas de la distribución). Por lo tanto, a las pruebas de este tipo se les llama *pruebas bilaterales* o *pruebas de dos colas*.

Sin embargo, hay ocasiones en las que interesan únicamente los valores extremos a un solo lado de la media (es decir, en una sola cola de la distribución); por ejemplo, cuando se prueba si un método es mejor que otro (que es distinto a probar si un método es mejor o peor que otro). A este tipo de pruebas se les llama *pruebas unilaterales* o *pruebas de una cola*. En estos casos la región crítica es una región en un solo lado de la distribución y su área es igual al nivel de significancia.

La tabla 10.1, en la que se dan los valores críticos de z tanto para pruebas de una cola como para pruebas de dos colas correspondientes a varios niveles de significancia, se encontrará útil como referencia. Valores críticos de z para otros niveles de significancia se encuentran en la tabla de áreas de la curva normal (apéndice II).

PRUEBAS ESPECIALES

Cuando las muestras son grandes, las distribuciones muestrales de muchos estadísticos tienen una distribución normal (o por lo menos aproximadamente normal), y en estas pruebas se puede emplear la correspondiente puntuación z. Los siguientes casos especiales, tomados de la tabla 8.1, son sólo algunos de los estadísticos de interés práctico. En cada uno de estos casos, el resultado es válido para poblaciones infinitas o cuando el muestreo se hace con reposición. Si el muestreo se hace de poblaciones finitas y sin reposición, es necesario modificar las fórmulas. Ver la página 182.

1. **Media.** Aquí $S = \bar{X}$, la media muestral; $\mu_S = \mu_{\bar{X}} = \mu$, la media poblacional, y $\sigma_S = \sigma_{\bar{X}} = \sigma/\sqrt{N}$, donde σ es la desviación estándar poblacional y N es el tamaño de la muestra. La puntuación z está dada por

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

   Si es necesario, para estimar σ se emplea la desviación muestral s o $\hat{s}$.

2. **Proporciones.** Aquí $S = P$, la proporción de "éxitos" en una muestra; $\mu_S = \mu_P = p$, donde p es la proporción poblacional de éxitos y N es el tamaño de la muestra, y $\sigma_S = \sigma_P = \sqrt{pq/N}$, donde $q = 1 - p$.
   La puntuación z está dada por

$$z = \frac{P - p}{\sqrt{pq/N}}$$

   En el caso de $P = X/N$, donde $X =$ cantidad de éxitos obtenidos realmente en una muestra, la puntuación z se transforma en

$$z = \frac{X - Np}{\sqrt{Npq}}$$

   Es decir, $\mu_X = \mu = Np$, $\sigma_X = \sigma = \sqrt{Npq}$ y $S = X$.

Las fórmulas para otros estadísticos se pueden obtener de manera similar.

CURVA CARACTERÍSTICA DE OPERACIÓN; POTENCIA DE UNA PRUEBA

Se ha visto cómo limitar el error tipo I eligiendo de manera adecuada el nivel de significancia. Para evitar totalmente cometer un error tipo II, simplemente no hay que cometerlo, que es lo mismo que no aceptar ninguna hipótesis. Sin embargo, en la práctica esto no es posible. Entonces lo que se hace es emplear las *curvas características de operación* o *curvas OC*, que son curvas que muestran la probabilidad de cometer un error tipo II bajo diversas hipótesis. Estas curvas proporcionan indicaciones de qué tan bien permite una prueba determinada minimizar los errores tipo II; es decir, indican la *potencia de una prueba* para evitar que se cometan errores de decisión. Estas curvas son útiles en el diseño de experimentos, ya que muestran informaciones como qué tamaño de muestra emplear.

VALOR p EN PRUEBAS DE HIPÓTESIS

El valor p es la probabilidad de obtener un estadístico muestral tan extremo o más extremo que el obtenido, suponiendo que la hipótesis nula sea verdadera. Para probar una hipótesis empleando este método se establece un valor α; se

calcula el valor p y si el valor $p \leq \alpha$, se rechaza H_0. En caso contrario, no se rechaza H_0. En pruebas para medias empleando muestras grandes ($n > 30$), el valor p se calcula como sigue:

1. Para $H_0 : \mu = \mu_0$ y $H_1 : \mu < \mu_0$, valor $p = P(Z <$ el estadístico de prueba calculado).
2. Para $H_0 : \mu = \mu_0$ y $H_1 : \mu > \mu_0$, valor $p = P(Z >$ el estadístico de prueba calculado).
3. Para $H_0 : \mu = \mu_0$ y $H_1 : \mu \neq \mu_0$, valor $p = P(Z < -|$ el estadístico de prueba calculado $|) + P(Z > |$ el estadístico de prueba calculado $|)$.

El estadístico de prueba calculado es $\dfrac{\bar{x} - \mu_0}{(s/\sqrt{n})}$, donde $\bar{x}$ es la media de la muestra, s es la desviación estándar de la muestra y μ_0 es el valor que se ha especificado para μ en la hipótesis nula. Obsérvese que σ no se conoce, se estima a partir de la muestra y se usa s. Este método para pruebas de hipótesis es equivalente al método de hallar el o los valores críticos y si el estadístico de prueba cae en la región de rechazo, rechazar la hipótesis nula. Usando cualquiera de estos métodos se llega a la misma decisión.

GRÁFICAS DE CONTROL

En la práctica suele ser importante darse cuenta cuándo un proceso ha cambiado lo suficiente como para que se deban tomar medidas para remediar la situación. Estos problemas surgen, por ejemplo, en el control de calidad. Los supervisores de control de calidad deben decidir si los cambios observados se deben sólo a fluctuaciones casuales o a verdaderos cambios en el proceso de fabricación debidos al deterioro de las máquinas, a los empleados, a errores, etc. Las *gráficas de control* proporcionan un método útil y sencillo para tratar tales problemas (ver problema 10.16).

PRUEBAS PARA DIFERENCIAS MUESTRALES

Diferencias entre medias

Sean $\bar{X}_1$ y $\bar{X}_2$ las medias muestrales de muestras grandes de tamaños N_1 y N_2 obtenidas de poblaciones cuyas medias son μ_1 y μ_2 y cuyas desviaciones estándar son σ_1 y σ_2, respectivamente. Considérese la hipótesis nula de que *no hay diferencia* entre las dos medias poblacionales (es decir, $\mu_1 = \mu_2$), lo cual es equivalente a decir que las muestras se han tomado de dos poblaciones que tienen la misma media.

Haciendo $\mu_1 = \mu_2$ en la ecuación (*5*) del capítulo 8 se ve que la distribución muestral de las diferencias entre las medias es aproximadamente normal con media y desviación estándar dadas por

$$\mu_{\bar{X}1-\bar{X}2} = 0 \qquad \text{y} \qquad \sigma_{\bar{X}1-\bar{X}2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} \tag{1}$$

donde, si es necesario, se pueden usar las desviaciones estándar muestrales s_1 y s_2 (o $\hat{s}_1$ y $\hat{s}_2$) como estimaciones de σ_1 y σ_2.

Empleando la variable estandarizada, o puntuación z, dada por

$$z = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sigma_{\bar{X}1-\bar{X}2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}1-\bar{X}2}} \tag{2}$$

se puede probar la hipótesis nula contra la hipótesis alternativa (o la significancia de la diferencia observada) a un nivel de significancia apropiado.

Diferencias entre proporciones

Sean P_1 y P_2 las proporciones muestrales de muestras grandes de tamaños N_1 y N_2 obtenidas de poblaciones cuyas proporciones son p_1 y p_2. Considérese la hipótesis nula de que *no hay diferencia* entre estos parámetros poblacionales (es decir, $p_1 = p_2$) y que por lo tanto las muestras se han obtenido realmente de la misma población.

Haciendo, en la ecuación (6) del capítulo 8, $p_1 = p_2 = p$, se ve que la distribución muestral de las diferencias entre las proporciones es aproximadamente normal, y que su media y su desviación estándar están dadas por

$$\mu_{P1}-\mu_{P2} = 0 \qquad \text{y} \qquad \sigma_{P1-P2} = \sqrt{pq\left(\frac{1}{N_1}+\frac{1}{N_2}\right)} \tag{3}$$

donde

$$p = \frac{N_1P_1 + N_2P_2}{N_1 + N_2}$$

se usa como estimación de la proporción poblacional y donde $q = 1 - p$.

Empleando la variable estandarizada

$$z = \frac{P_1 - P_2 - 0}{\sigma_{P1-P2}} = \frac{P_1 - P_2}{\sigma_{P1-P2}} \tag{4}$$

se puede probar la diferencia observada a nivel de significancia apropiado y con esto probar la hipótesis nula.

Se pueden hacer pruebas con otros estadísticos de manera similar.

PRUEBAS EMPLEANDO DISTRIBUCIONES BINOMIALES

Las pruebas en las que se usen distribuciones binomiales (así como otras distribuciones) pueden hacerse de manera análoga a las pruebas en las que se emplean distribuciones normales; el principio básico es esencialmente el mismo. Ver los problemas del 10.23 al 10.28.

PROBLEMAS RESUELTOS

PRUEBAS DE MEDIAS Y PROPORCIONES EMPLEANDO DISTRIBUCIONES NORMALES

10.1 Encontrar la probabilidad de obtener entre 40 y 60 caras inclusive en 100 lanzamientos de una moneda que no esté cargada.

SOLUCIÓN

De acuerdo con la probabilidad binomial, la probabilidad buscada es

$$\binom{100}{40}\left(\frac{1}{2}\right)^{40}\left(\frac{1}{2}\right)^{60} + \binom{100}{41}\left(\frac{1}{2}\right)^{41}\left(\frac{1}{2}\right)^{59} + \cdots + \binom{100}{60}\left(\frac{1}{2}\right)^{60}\left(\frac{1}{2}\right)^{40}$$

Como tanto $Np = 100(\frac{1}{2})$ como $Nq = 100(\frac{1}{2})$ son mayores que 5, para evaluar esta suma puede emplearse la aproximación normal a la distribución binomial. La media y la desviación estándar de la cantidad de caras en 100 lanzamientos están dadas por

$$\mu = Np = 100(\tfrac{1}{2}) = 50 \qquad \text{y} \qquad \sigma = \sqrt{Npq} = \sqrt{(100)(\tfrac{1}{2})(\tfrac{1}{2})} = 5$$

En una escala continua, entre 40 y 60 caras corresponden a entre 39.5 y 60.5 caras. Por lo tanto, se tiene

$$39.5 \text{ en unidades estándar} = \frac{39.5 - 50}{5} = -2.10 \qquad 60.5 \text{ en unidades estándar} = \frac{60.5 - 50}{5} = 2.10$$

$$\begin{aligned}\text{Probabilidad buscada} &= \text{área bajo la curva normal entre } z = -2.10 \text{ y } z = 2.10 \\ &= 2(\text{área entre } z = 0 \text{ y } z = 2.10) = 2(0.4821) = 0.9642.\end{aligned}$$

10.2 Para probar la hipótesis de que una moneda no está cargada se adopta la siguiente regla de decisión:

Aceptar la hipótesis si el número de caras de una sola muestra de 100 lanzamientos está entre 40 y 60 inclusive.
Rechazar la hipótesis si no es así.

a) Encontrar la probabilidad de rechazar la hipótesis en caso de que en realidad sea correcta.
b) Graficar la regla de decisión y el resultado del inciso *a*).
c) ¿A qué conclusión se llega si en la muestra de 100 lanzamientos se obtienen 53 caras? ¿Y si se obtienen 60 caras?
d) ¿Puede estar equivocada la conclusión obtenida en el inciso *c*)?

SOLUCIÓN

a) De acuerdo con el problema 10.1, la probabilidad de que no se obtengan entre 40 y 60 caras inclusive si la moneda no está cargada es $1 - 0.9642 = 0.0358$. Por lo tanto, la probabilidad de rechazar la hipótesis (nula) cuando en realidad sea correcta es 0.0358.

b) En la figura 10.2 se ilustra la regla de decisión. Se muestra la distribución de probabilidad para la obtención de caras en 100 lanzamientos de una moneda no cargada. Si en una sola muestra de 100 lanzamientos se obtiene una puntuación z entre -2.10 y 2.10, se acepta la hipótesis; si no es así, se rechaza la hipótesis y se concluye que la moneda está cargada.

El error que se comete si se rechaza la hipótesis cuando en realidad deba aceptarse es el *error tipo I* de la regla de decisión, y la probabilidad de cometer este error es igual a 0.0358, de acuerdo con el inciso *a*); este error está representado por el total del área sombreada de la figura. Si en una sola muestra de 100 lanzamientos se obtiene una cantidad de caras cuya puntuación z (o estadístico z) se encuentra en la región sombreada, se dice que la puntuación z difiere de manera *significativa* de lo que se esperaría si la hipótesis fuera verdadera. Es por esta razón que a la región sombreada (es decir, a la probabilidad de cometer un error tipo I) se le conoce como *nivel de significancia* de la regla de decisión, que en este caso es igual a 0.0358. Por lo tanto, se habla de rechazo de la hipótesis a nivel de significancia 0.0358 (o 3.58%).

c) De acuerdo con la regla de decisión, en ambos casos debe aceptarse la hipótesis de que la moneda no está cargada. Puede argumentarse que bastará que se obtenga una cara más para que se rechace la hipótesis. Esto es a lo que se enfrenta cuando se emplea una clara línea divisoria para tomar una decisión.

d) Sí. Tal vez se acepte la hipótesis cuando en realidad debería haberse rechazado, que sería el caso, por ejemplo, si la probabilidad de cara fuera en realidad 0.7 en lugar de 0.5. El error que se comete al aceptar una hipótesis que debería rechazarse es un *error tipo II* de la decisión.

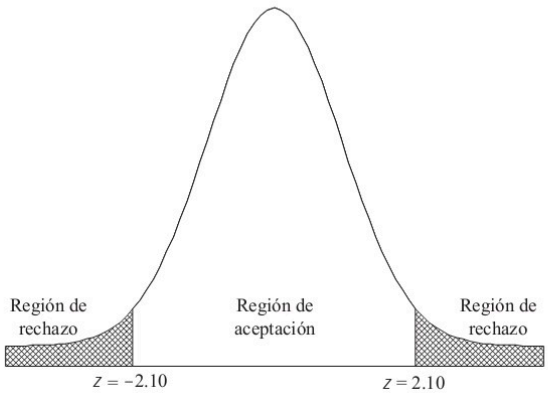


Figura 10-2 Curva normal estándar en la que se muestran las regiones de aceptación y de rechazo para probar que una moneda no está cargada.

10.3 Empleando la distribución binomial y no la aproximación normal a la distribución binomial, diseñar una regla de decisión para probar la hipótesis de que una moneda no está cargada si se emplea una muestra de 64 lanzamientos y se usa como nivel de significancia 0.05. Usar MINITAB como ayuda para encontrar la solución.

SOLUCIÓN

En la figura 10-3 se presenta la gráfica de probabilidades binomiales cuando una moneda no cargada se lanza 64 veces. Abajo de la figura 10-3 se presentan las probabilidades acumuladas generadas con MINITAB.

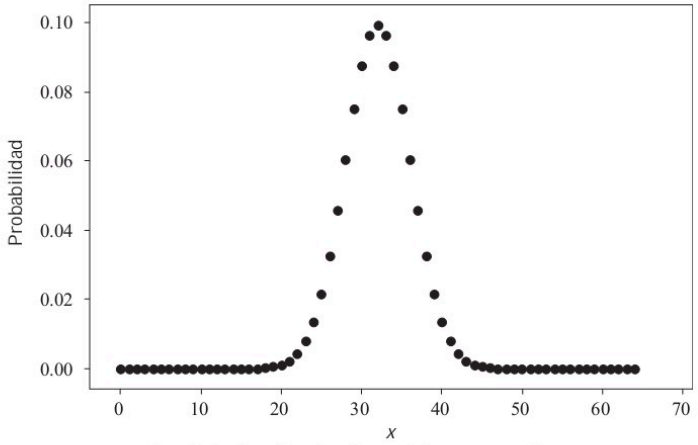


Figura 10-3 MINITAB, gráfica de la distribución binomial correspondiente a $n = 64$ y $p = 0.5$.

x	Probabilidad	Acumulada	x	Probabilidad	Acumulada
0	0.0000000	0.0000000	13	0.0000007	0.0000009
1	0.0000000	0.0000000	14	0.0000026	0.0000035
2	0.0000000	0.0000000	15	0.0000086	0.0000122
3	0.0000000	0.0000000	16	0.0000265	0.0000387
4	0.0000000	0.0000000	17	0.0000748	0.0001134
5	0.0000000	0.0000000	18	0.0001952	0.0003087
6	0.0000000	0.0000000	19	0.0004727	0.0007814
7	0.0000000	0.0000000	20	0.0010636	0.0018450
8	0.0000000	0.0000000	21	0.0022285	0.0040735
9	0.0000000	0.0000000	22	0.0043556	0.0084291
10	0.0000000	0.0000000	23	0.0079538	0.0163829
11	0.0000000	0.0000001	24	0.0135877	0.0299706
12	0.0000002	0.0000002	25	0.0217403	0.0517109

Como se ve, $P(X \leq 23) = 0.01638$. Como la distribución es simétrica, se sabe también que $P(X \geq 41) = 0.01638$. La región de rechazo $\{X \leq 23$ y $X \geq 41\}$ tiene la probabilidad $2(0.01638) = 0.03276$. La región de rechazo $\{X \leq 24$ y $X \geq 40\}$ es mayor que 0.05. Cuando se usa una distribución binomial no se puede tener una región de rechazo exactamente igual a 0.05. Lo más cercano a 0.05 que se puede tener, sin que se tenga una probabilidad mayor a este valor, es 0.03276.

Resumiendo, la moneda se lanza 64 veces. Se declarará que está cargada, o no equilibrada, si se obtienen 23 o menos, o 41 o más caras. La posibilidad de cometer un error tipo I es 0.03276, que es lo más cerca que se puede estar de 0.05, sin sobrepasar este valor.

10.4 Volver al problema 10.3. Usando la distribución binomial, no la aproximación normal a la distribución binomial, diseñar una regla de decisión para probar la hipótesis de que la moneda no está cargada empleando una mues-

tra de 64 lanzamientos de la moneda y un nivel de significancia de 0.05. Emplear EXCEL como ayuda para dar la solución.

SOLUCIÓN

En la columna A de la hoja de cálculo de EXCEL se ingresan los resultados 0 a 64. Las expresiones =BINOMDIST(A1,64,0.5,0) y =BINOMDIST(A1,64,0.5,1) se emplean para obtener la distribución binomial y la distribución binomial acumulada. El 0, que aparece como cuarto parámetro, indica que se requieren probabilidades individuales, y el 1 indica que se desean las probabilidades acumuladas. Haciendo clic y arrastrando en la columna B se obtienen las probabilidades individuales y haciendo clic y arrastrando en la columna C se obtienen las probabilidades acumuladas.

A	B	C	A	B	C
x	Probabilidad	Acumulada	x	Probabilidad	Acumulada
0	5.42101E-20	5.42101E-20	13	7.12151E-07	9.40481E-07
1	3.46945E-18	3.52366E-18	14	2.59426E-06	3.53474E-06
2	1.09288E-16	1.12811E-16	15	8.64754E-06	1.21823E-05
3	2.25861E-15	2.37142E-15	16	2.64831E-05	3.86654E-05
4	3.44438E-14	3.68152E-14	17	7.47758E-05	0.000113441
5	4.13326E-13	4.50141E-13	18	0.000195248	0.000308689
6	4.06437E-12	4.51451E-12	19	0.000472706	0.000781395
7	3.36762E-11	3.81907E-11	20	0.001063587	0.001844982
8	2.39943E-10	2.78134E-10	21	0.002228469	0.004073451
9	1.49298E-09	1.77111E-09	22	0.004355644	0.008429095
10	8.21138E-09	9.98249E-09	23	0.007953785	0.01638288
11	4.03104E-08	5.02929E-08	24	0.013587715	0.029970595
12	1.78038E-07	2.28331E-07	25	0.021740344	0.051710939

Se encuentra, como en el problema 10.3, que $P(X \leq 23) = 0.01638$ y debido a la simetría, $P(X \geq 41) = 0.01638$, y que la región de rechazo es $\{X \leq 23$ o $X \geq 41\}$ y el nivel de significancia es $0.01638 + 0.01638$ o bien 0.03276.

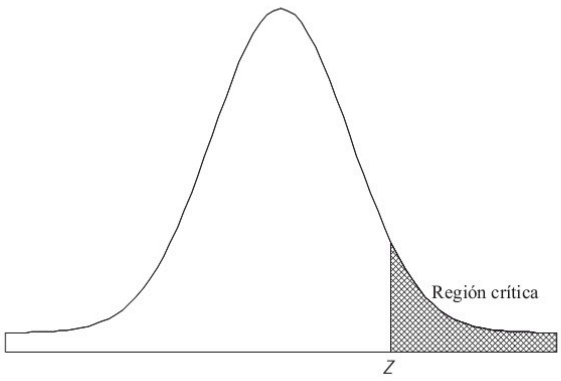


Figura 10-4 Determinación del valor Z que dará una región crítica igual a 0.05.

10.5 Se realiza un experimento de percepción extrasensorial (PES) en el que se pide a un individuo que está en una habitación que adivine el color (rojo o verde) de una carta extraída de un juego de 50 cartas bien mezcladas por una persona en otra habitación. El individuo no sabe cuántas cartas rojas o verdes hay en ese conjunto de cartas. Si este individuo identifica 32 cartas correctamente, determinar si los resultados son significativos al nivel: *a*) 0.05 y *b*) 0.01.

SOLUCIÓN

Si p es la probabilidad de que la persona elija correctamente el color de la carta, entonces hay que decidir entre las dos hipótesis:

H_0: $p = 0.5$, el individuo simplemente está adivinando (es decir, el resultado se debe a la casualidad).
H_1: $p > 0.5$, la persona tiene PES.

Como lo que interesa no es la habilidad de la persona para obtener puntuaciones extremadamente bajas, sino sólo su habilidad para obtener puntuaciones altas, se elige una prueba de una cola. Si la hipótesis H_0 es verdadera, entonces la media y la desviación estándar de la cantidad de cartas identificadas correctamente están dadas por

$$\mu = Np = 50(0.5) = 25 \qquad \text{y} \qquad \sigma = \sqrt{Npq} = \sqrt{50(0.5)(0.5)} = \sqrt{12.5} = 3.54$$

a) Como se trata de una prueba de una cola con nivel de significancia 0.05, en la figura 10-4 se debe elegir z de manera que el área sombreada, en la región crítica de puntuaciones altas, sea 0.05. El área entre 0 y z será 0.4500 y $z = 1.645$; este valor también se puede leer en la tabla 10.1. Por lo tanto, la regla de decisión (o prueba de significancia) es:

Si la puntuación z observada es mayor a 1.645, los resultados son significativos al nivel 0.05 y la persona tiene poderes extrasensoriales.

Si la puntuación z es menor a 1.645, los resultados se deben a la casualidad (es decir, no son significativos al nivel 0.05).

Como 32 en unidades estándar $(32 - 25)/3.54 = 1.98$, lo cual es mayor a 1.645, se concluye, al nivel 0.05, que la persona tiene poderes extrasensoriales.

Obsérvese que en realidad debe aplicarse la corrección por continuidad, ya que 32 en una escala continua está entre 31.5 y 32.5. Sin embargo, la puntuación estándar correspondiente a 31.5 es $(31.5 - 25)/3.54 = 1.84$, con lo que se llega a la misma conclusión.

b) Si el nivel de significancia es 0.01, entonces el área entre 0 y z es 0.4900, de donde se concluye que $z = 2.33$.

Como 32 (o 31.5) en unidades estándar es 1.98 (o 1.84), que es menor a 2.33, se concluye que los resultados *no son significativos* al nivel 0.01.

Algunos especialistas en estadística adoptan la siguiente terminología: resultados significativos al nivel 0.01 son *altamente significativos*; resultados significativos al nivel 0.05, pero no al nivel 0.01, son *probablemente significativos*, y resultados significativos a niveles mayores a 0.05 *no son significativos*. De acuerdo con esta terminología se concluye que los resultados experimentales anteriores son *probablemente significativos*, de manera que será necesario hacer más investigaciones acerca del fenómeno.

Como los niveles de significancia sirven de guía en la toma de decisiones, algunos especialistas en estadística dan las probabilidades empleadas. Por ejemplo, como en este problema, $\Pr\{z \geq 1.84\} = 0.0322$, un especialista en estadística dirá que con base en el experimento, las posibilidades de estar equivocado al concluir que la persona tiene poderes extrasensoriales son aproximadamente 3 en 100. A la probabilidad que se da (0.0322 en este caso) se le conoce como valor p de la prueba.

10.6 Se asegura que 40% de las personas que hacen sus declaraciones de impuestos, las hacen empleando algún software para impuestos. En una muestra de 50 personas, 14 emplearon software para hacer su declaración de impuestos. Probar H_0: $p = 0.4$ versus H_a: $p < 0.4$ a $\alpha = 0.05$, donde p es la proporción poblacional de los que emplean software para hacer su declaración de impuestos. Haga la prueba empleando la distribución binomial y también empleando la aproximación normal a la distribución binomial.

SOLUCIÓN

Si se emplea la prueba exacta H_0: $p = 0.4$ versus H_a: $p < 0.4$ a $\alpha = 0.05$, la hipótesis nula se rechaza si $X \leq 15$. A esta región se le llama la región de rechazo. Si se emplea la prueba basada en la aproximación normal a la binomial, la hipótesis nula se rechaza si $Z < -1.645$ y a esta región se le llama la región de rechazo. A $X = 14$ se le llama estadístico de prueba. El estadístico de prueba binomial está en la región de rechazo y la hipótesis nula se rechaza. Usando la aproximación normal, el estadístico de prueba es $z = \frac{14-20}{3.46} = -1.73$. El verdadero valor de α es 0.054 y la región de rechazo es $X \leq 15$ y se emplea la probabilidad binomial acumulada $P(X \leq 15)$. Empleando la aproximación normal también se rechazará la hipótesis nula, ya que $z = -1.73$ está en la región de rechazo que es $Z < -1.645$. Obsérvese que si se usa la distribución binomial para realizar la prueba, el estadístico de prueba tiene una distribución binomial. Si se emplea la distribución normal, el estadístico de prueba, Z, tiene una distribución normal estándar.

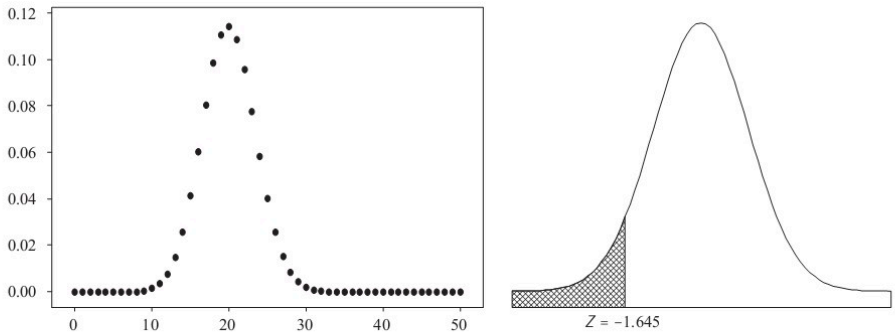


Figura 10-5 Comparación entre la prueba exacta a la izquierda (binomial) y la prueba aproximada a la derecha (normal estándar).

10.7 El valor p en una prueba de hipótesis se define como el menor nivel de significancia al cual se rechaza la hipótesis nula. En este problema se ilustra el cálculo del valor p para un estadístico de prueba. Usar los datos del problema 9.6 para probar la hipótesis nula de que la altura media de los árboles es igual a 5 pies (ft) contra la hipótesis alternativa de que la altura media es menor a 5 ft. Encontrar el valor p de esta prueba.

SOLUCIÓN

El valor encontrado para z es $z = (59.22 - 60)/1.01 = -0.77$. El menor nivel de significancia al que se rechaza la hipótesis nula es el valor $p = P(z < -0.77) = 0.5 - 0.2794 = 0.2206$. La hipótesis nula se rechaza si el valor p es menor al nivel de significancia preestablecido. En este problema, si el nivel de significancia preestablecido es 0.05, no se rechaza la hipótesis nula. A continuación se presenta la solución que da MINITAB, donde el comando `Alternative-1` indica que se trata de una prueba de la cola inferior.

```
MTB> ZTest mean = 60 sd = 10.111 data in cl ;
SUBC>Alternative -1
```

Prueba Z

```
Test of mu = 60.00 vs mu < 60.00
The assumed sigma = 10.1

Variable      N     Mean     StDev     SE Mean         Z         P
height      100    59.22     10.11        1.01     -0.77      0.22
```

10.8 Se toma una muestra de 33 personas que escuchen radio y se determina la cantidad de horas, por semana, que escuchan la radio. Los datos son los siguientes.

9 8 7 4 8 6 8 8 7 10 8 10 6 7 7 8 9
6 5 8 5 6 8 7 8 5 5 8 7 6 6 4 5

Probar, de las siguientes tres maneras equivalentes, la hipótesis nula $\mu = 5$ horas (h) contra la hipótesis alternativa $\mu \neq 5$ h al nivel de significancia $\alpha = 0.05$:

a) Calcular el valor del estadístico de prueba y compararlo con el valor crítico correspondiente a $\alpha = 0.05$.
b) Calcular el valor p del estadístico de prueba encontrado y comparar este valor p con $\alpha = 0.05$.
c) Calcular el intervalo de confianza $1 - \alpha = 0.95$ para μ y determinar si 5 cae dentro de este intervalo.

SOLUCIÓN

En el siguiente resultado de MINITAB se halla, primero, la desviación estándar y después se emplea en las declaraciones `Ztest` y `Zinterval`.

```
MTB > standard deviation c1
Standard deviation of hours = 1.6005

MTB > ZTest 5.01.6005 'hours' ;
SUBC> Alternative 0.
```

Prueba Z

```
Test of mu = 5.000 vs mu not = 5.000
The assumed sigma = 1.60

Variable      N      Mean     StDev     SE Mean         Z          P
hours        33     6.897     1.600       0.279      6.81     0.0000

MTB > ZInterval 95.01.6005    'hours'
Variable      N      Mean     StDev     SE Mean          95.0 % CI
hours        33     6.897     1.600       0.279     ( 6.351, 7.443)
```

a) El valor calculado para el estadístico de prueba es $Z = \dfrac{6.897 - 5}{0.279} = 6.81$, los valores críticos son ± 1.96, y la hipótesis nula se rechaza. Obsérvese que éste es el valor encontrado que aparece en el resultado de MINITAB.

b) El valor p encontrado, de acuerdo con los resultados de MINITAB, es 0.0000, por lo tanto, el valor $p < \alpha = 0.05$, la hipótesis nula se rechaza.

c) Como el valor especificado por la hipótesis nula, 5, no está contenido en el intervalo de confianza de 95% para μ, la hipótesis nula se rechaza.

Estos tres procedimientos para probar una hipótesis nula contra una de hipótesis alternativa de dos colas son equivalentes.

10.9 La resistencia a la ruptura de los cables fabricados por una empresa tiene media de 1 800 libras (lb) y desviación estándar de 100 lb. Se asegura que mediante una nueva técnica puede aumentarse la resistencia a la ruptura. Para probar esto, se prueba una muestra de 50 cables y se encuentra que su resistencia media a la ruptura es 1 850 lb. ¿Puede apoyarse, a nivel de significancia 0.01, la aseveración hecha antes?

SOLUCIÓN

Se tiene que decidir entre las dos hipótesis siguientes:

H_0: $\mu = 1\,800$ lb, en realidad no hay cambio en la resistencia a la ruptura.
H_1: $\mu > 1\,800$ lb, sí hay cambio en la resistencia a la ruptura.

Por lo tanto, se debe usar una prueba de una cola; el diagrama correspondiente a esta prueba es idéntico al de la figura 10-4 del problema 10.5*a*). A nivel de significancia 0.01, la regla de decisión es:

Si la puntuación z observada es mayor a 2.33, los resultados son significativos a nivel de significancia 0.01 y H_0 se rechaza.
Si no es así, H_0 se acepta (o la decisión se aplaza).

Bajo la hipótesis de que H_0 es verdadera, se encuentra que

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} = \frac{1\,850 - 1\,800}{100/\sqrt{50}} = 3.55$$

que es mayor a 2.33. Por lo tanto, se concluye que los resultados son *altamente significativos* y que la aseveración hecha puede apoyarse.

VALORES *p* PARA PRUEBAS DE HIPÓTESIS

10.10 A un grupo de 50 compradores se le preguntó cuánto gastaba anualmente en sus compras por Internet. En la tabla 10.2 se muestran las respuestas. Se desea probar que gastan \$325 por año contra una cantidad diferente a \$325. Encontrar el valor *p* para la prueba de hipótesis. ¿A qué conclusión se llega empleando $\alpha = 0.05$?

Tabla 10.2

418	379	77	212	378
363	434	348	245	341
331	356	423	330	247
351	151	220	383	257
307	297	448	391	210
158	310	331	348	124
523	356	210	364	406
331	364	352	299	221
466	150	282	221	432
366	195	96	219	202

SOLUCIÓN

La media de estos datos es 304.60, la desviación estándar es 101.51, el estadístico de prueba obtenido es

$$z = \frac{304.60 - 325}{101.50/\sqrt{50}} = -1.43.$$

El estadístico Z tiene aproximadamente una distribución normal estándar. El valor p calculado es el siguiente $P(Z < -|\text{estadístico de prueba calculado}|)$ o $Z > |\text{estadístico de prueba calculado}|)$ o $P(Z < -1.43) + P(Z > 1.43)$. La respuesta puede hallarse usando el apéndice II o usando EXCEL. Mediante EXCEL, el valor p =2*NORMDIST(−1.43) = 0.1527, dado que la curva normal es simétrica y las áreas a la izquierda de −1.43 y a la derecha de 1.43 son iguales, se puede simplemente duplicar en el área a la izquierda de −1.43. Como el valor p es menor a 0.05, no se rechaza la hipótesis nula. En la figura 10.6 se muestra gráficamente el valor p calculado en este problema.

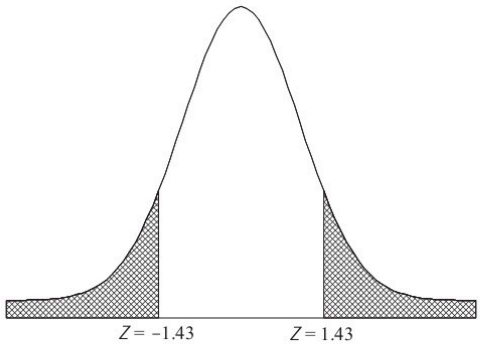


Figura 10-6 El valor *p* es la suma del área a la izquierda de *Z* = –1.43 más que el área a la derecha de *Z* = 1.43.

10.11 Volver al problema 10.10. Para analizar los datos usar el software para estadística de MINITAB. Obsérvese que el software da el valor p y al usuario se le deja la decisión respecto a la hipótesis de acuerdo con el valor que le haya asignado a α.

SOLUCIÓN

Con la secuencia **"Stat ⇒ Basic statistics ⇒ 1 sample Z"** se obtiene el análisis siguiente. El software calcula para el usuario el estadístico de prueba y el valor *p*.

Muestra uno Z: cantidad

```
Test of mu = 325 vs not = 325
The assumed standard deviation = 101.51

Variable      N      Mean     StDev     SE Mean          Z         P
Amount       50   304.460   101.508      14.356      -1.43     0.152
```

Obsérvese que el software proporciona el valor del estadístico de prueba (−1.43) y el valor *p* (0.152).

10.12 En la tabla 10.3 se muestran los resultados de un estudio sobre individuos que emplean la computadora para hacer sus declaraciones de impuestos. Los datos de la tabla dan el tiempo que necesitan para hacer su declaración. La hipótesis nula es $H_0 : \mu = 8.5$ horas contra la hipótesis alternativa, que es $H_1 : \mu < 8.5$. Encontrar el valor *p* de esta prueba de hipótesis. ¿A qué conclusión llega empleando $\alpha = 0.05$?

Tabla 10.3

6.2	4.8	8.9	5.6	6.5
11.5	8.6	6.2	8.5	5.2
2.7	14.9	11.2	6.9	7.9
4.8	9.5	12.4	9.7	10.7
8.0	11.8	7.4	9.1	4.9
9.1	6.4	9.5	7.6	6.7
2.6	3.5	6.4	4.3	7.9
3.3	10.3	3.2	11.5	1.7
10.4	8.5	10.8	6.9	5.3
4.9	4.4	9.4	5.6	7.0

SOLUCIÓN

La media de los datos que se presentan en la tabla 10.3 es 7.42 h, la desviación estándar es 2.91 h y el estadístico de prueba calculado es $Z = \dfrac{7.42 - 8.5}{2.91/\sqrt{50}} = -2.62$. El estadístico Z tiene aproximadamente la distribución normal estándar. Con la secuencia de MINITAB **"Calc ⇒ Probability distribution ⇒ Normal"** se obtiene el cuadro de diálogo que se muestra en la figura 10-7. El cuadro de diálogo se llena como se indica.

Los resultados que da el cuadro de diálogo de la figura 10-7 son los siguientes:

Función de distribución acumulada

```
Normal with mean = 0 and standard deviation = 1

x      P ( X<=x)
-2.62  0.0043965
```

El valor *p* es 0.0044 y como el valor $p < \alpha$, se rechaza la hipótesis nula. Consultar la figura 10-8 para ver gráficamente el valor *p* obtenido en este problema.

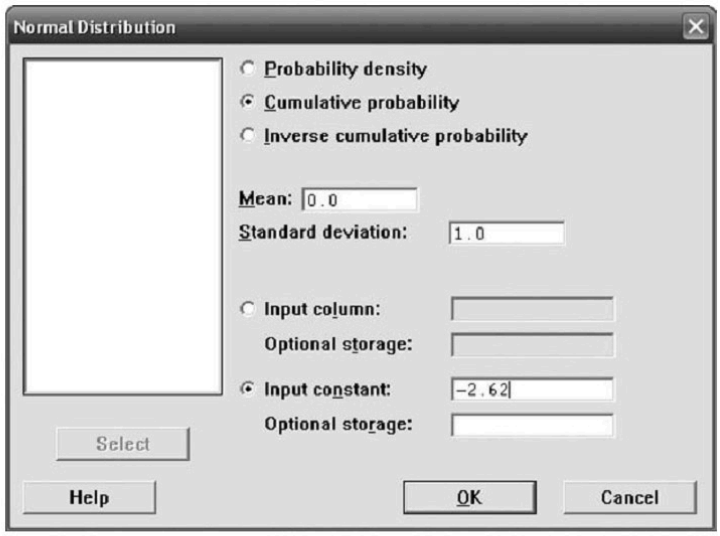


Figura 10-7 Cuadro de diálogo para calcular el valor p si el estadístico de prueba es igual a –2.62.

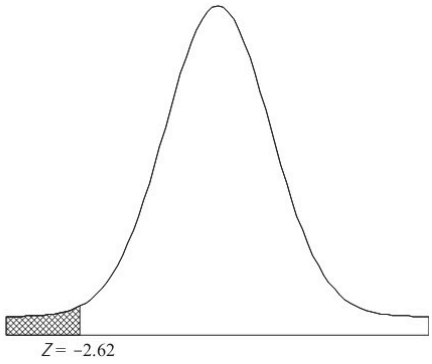


Figura 10-8 El valor p es el área a la izquierda de $Z = -2.62$.

10.13 Refiérase al problema 10.12. Para analizar los datos usar el software para estadística SAS. Obsérvese que este software da el valor p y deja al usuario la decisión respecto de la hipótesis de acuerdo con el valor que el usuario haya asignado a α.

SOLUCIÓN

A continuación se presentan los resultados dados por SAS. El valor p se da como Prob $> z = 0.0044$, el mismo valor que se obtuvo en el problema 10.12. Este valor es el área bajo la curva normal estándar a la izquierda de -2.62. Comparar las demás cantidades dadas como resultados de SAS con las del problema 10.12.

RESULTADOS DE SAS

```
One Sample Z Test for a Mean
Sample Statistics for time
          N           Mean          Std. Dev    Std Error
-------------------------------------------------------------
         50           7.42            2.91         0.41
```

```
Hypothesis Test
Null hypothesis      Mean of time => 8.5
Alternative          Mean of time <  8.5
with a specified know standard deviation of 2.91
              Z Statistic     Prob > Z
              -----------------------
                -2.619        0.0044
95% Confidence Interval for the Mean
(Upper Bound Only)
              Lower Limit     Upper Limit
              -----------------------
               -infinity      8.10
```

Obsérvese que el intervalo unilateral de 95% $(-\infty, 8.10)$ no contiene el valor de la hipótesis nula, 8.5. Ésta es otra indicación de que la hipótesis nula se debe rechazar a nivel $\alpha = 0.05$.

10.14 Se asegura que el promedio de tiempo que escuchan MP3 las personas que utilizan estos dispositivos es 5.5 h por semana, contra un promedio mayor a 5.5. En la tabla 10.4 se dan las cantidades de tiempo que 50 personas pasan escuchando un MP3. Probar $H_0: \mu = 5.5$ h contra la hipótesis alternativa $H_1: \mu > 5.5$ h. Encontrar el valor p de esta prueba de hipótesis usando STATISTIX. ¿A qué conclusión se llega empleando $\alpha = 0.05$?

Tabla 10.4

6.4	6.4	6.8	7.6	6.9
5.8	5.9	6.9	5.9	6.0
6.3	5.5	6.1	6.4	4.8
6.3	4.2	6.2	5.0	5.9
6.5	6.8	6.8	5.1	6.5
6.7	5.4	5.9	3.5	4.4
6.9	6.7	6.4	5.1	5.4
4.7	7.0	6.0	5.8	5.8
5.7	5.2	4.9	6.6	8.2
6.9	5.5	5.2	3.3	8.3

SOLUCIÓN

STATISTIX proporciona los resultados siguientes:

```
Statistix 8.0
Descriptive Statistics
Variable       N           Mean              SD
MP3            50        5.9700          1.0158
```

El estadístico de prueba calculado es $Z = \dfrac{5.97 - 5.5}{1.0158/\sqrt{50}} = 3.27$. En la figura 10-9 se calcula el valor p.
En la figura 10-10 se muestra gráficamente el valor p encontrado en este problema.
El valor p encontrado es 0.00054 y como es menor a 0.05, se rechaza la hipótesis nula.

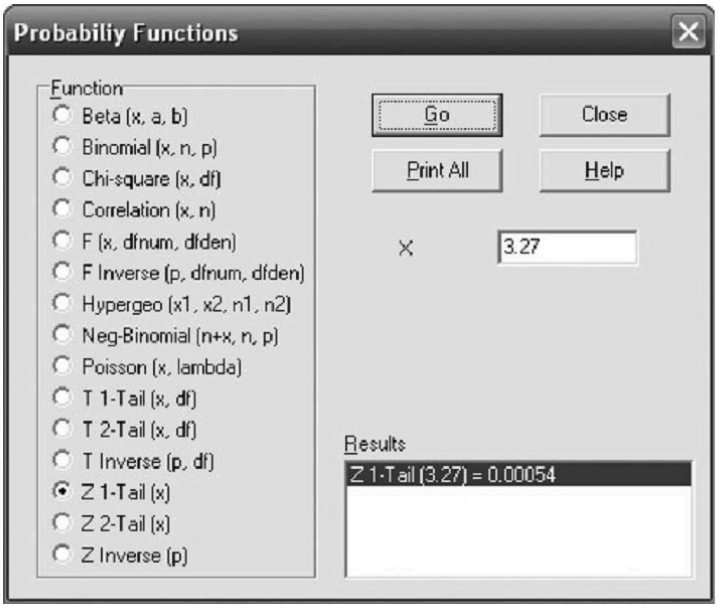


Figura 10-9 Cuadro de diálogo para hallar el valor p siendo el estadístico de prueba igual a 3.27.

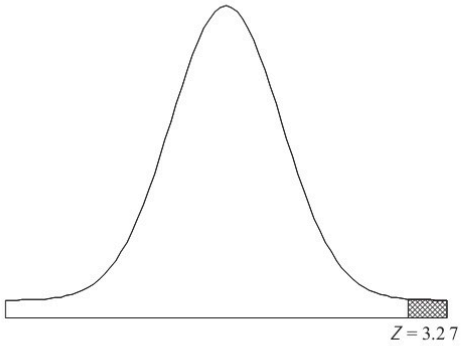


Figura 10-10 El valor p es el área a la derecha de $z = 3.27$.

10.15 Empleando SPSS, usar la secuencia **"Analyze ⇒ Compare means ⇒ one-sample t test"** y los datos del problema 10.14 para probar $H_0: \mu = 5.5$ h contra la hipótesis alternativa $H_a: \mu > 5.5$ h a $\alpha = 0.05$ hallando el valor p y comparándolo con α.

SOLUCIÓN

Los resultados de SPSS son los siguientes:

One-sample statistics

	N	Media	Desv. estándar	Media error estándar
MPE	50	5.9700	1.0184	.14366

One-sample test

	Valor de prueba = 5.5					
	t	gl	Sigma (2 colas)	Diferencia media	Intervalo de confianza de 95% de la diferencia	
					Inferior	Superior
MPE	3.272	49	0.002	.47000	.1813	.7587

En la primera parte de los resultados de SPSS se dan los estadísticos necesarios. Obsérvese que al estadístico de prueba encontrado se le llama t y no z. Esto se debe a que para $n > 30$, la distribución t y la distribución z son muy similares. La distribución t tiene un parámetro llamado grados de libertad que es igual a $n - 1$. El valor p encontrado por SPSS es siempre un valor p para dos colas y se le conoce como sigma(2 colas). Este valor es igual a 0.002. El valor para 1 cola es $0.002/2 = 0.001$. Este valor es un valor cercano al encontrado en el problema 10.14 que es igual a 0.00054. Cuando se usa un software, el usuario debe estar atento a la idiosincrasia de ese software.

GRÁFICAS DE CONTROL

10.16 Para controlar el llenado de recipientes de mostaza se emplea una gráfica de control. La cantidad media de llenado es 496 gramos (g) y la desviación estándar es 5 g. Para determinar si la máquina llenadora está trabajando en forma adecuada, cada hora, a lo largo de las 8 h del día, se toma una muestra de cinco recipientes. En la tabla 10.5 se presentan los datos de dos días.

a) Diseñar una regla de decisión mediante la cual se pueda estar muy seguro de que la media de llenado se mantiene, durante estos dos días, en 496 g con una desviación estándar igual a 5 g.
b) Mostrar cómo graficar la regla de decisión del inciso *a*).

Tabla 10.5

1	2	3	4	5	6	7	8
492.2	486.2	493.6	508.6	503.4	494.9	497.5	490.5
487.9	489.5	503.2	497.8	493.4	492.3	497.0	503.0
493.8	495.9	486.0	493.4	493.9	502.9	493.8	496.4
495.4	494.1	498.4	495.8	493.8	502.8	497.1	489.7
491.7	494.0	496.5	508.0	501.3	498.9	488.3	492.6

9	10	11	12	13	14	15	16
492.2	486.2	493.6	508.6	503.4	494.9	497.5	490.5
487.9	489.5	503.2	497.8	493.4	492.3	497.0	503.0
493.8	495.9	486.0	493.4	493.9	502.9	493.8	496.4
495.4	494.1	498.4	495.8	493.8	502.8	497.1	489.7
491.7	494.0	496.5	508.0	501.3	498.9	488.3	492.6

SOLUCIÓN

a) Con una confianza de 99.73% puede decirse que la media muestral $\bar{x}$ debe encontrarse en el intervalo de $\mu_{\bar{x}} - 3\sigma_{\bar{x}}$ hasta $\mu_{\bar{x}} + 3\sigma_{\bar{x}}$ o bien, de: $\mu - 3\dfrac{\sigma}{\sqrt{n}}$ hasta: $\mu + 3\dfrac{\sigma}{\sqrt{n}}$. Como $\mu = 496$, $\sigma = 5$ y $n = 5$, se sigue que con una confianza

de 99.73% la media muestral debe estar en el intervalo de: $496 - 3\frac{5}{\sqrt{5}}$ hasta: $496 + 3\frac{5}{\sqrt{5}}$ o bien entre 489.29 y 502.71. Por lo tanto, la regla de decisión es la siguiente:

Si la media muestral cae dentro del intervalo de 489.29 g a 502.71 g, se supone que la máquina está llenando correctamente.

Si no es así, se concluye que la máquina de llenado no está trabajando en forma adecuada y se busca la razón por la que el llenado es incorrecto.

b) Empleando una gráfica como la de la figura 10-11, llamada *gráfica de control de calidad*, se puede llevar un registro de las medias muestrales. Cada vez que se calcula una media muestral se representa mediante un punto. Mientras estos puntos se encuentren entre el límite inferior y el límite superior, el proceso está bajo control. Si un punto se sale de estos límites de control puede ser que algo esté mal y se recomienda hacer una investigación.

Las 80 observaciones se ingresan en la columna C1. Con la secuencia **"Stat ⇒ Control Charts ⇒ Variable charts for subgroups ⇒ Xbar"** se abre la ventana de diálogo que, una vez llenada, da la gráfica de control que se muestra en la figura 10-11.

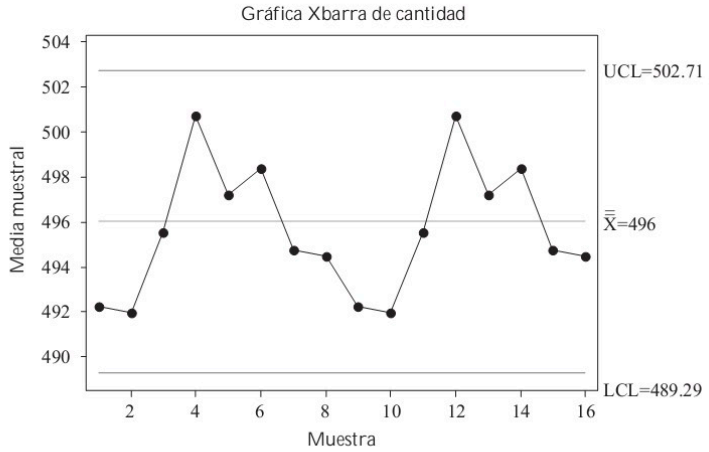


Figura 10-11 Gráfica de control con límites 3σ para el control de la media de llenado de los envases de mostaza.

Los límites de control especificados antes se conocen como límites de confianza del 99.73%, o simplemente, límites 3σ. También se pueden determinar otros límites de confianza (por ejemplo, límites del 99% o del 95%). En cada caso la elección depende de las circunstancias particulares.

PRUEBAS PARA DIFERENCIAS DE MEDIAS Y PROPORCIONES

10.17 A dos grupos de estudiantes, uno de 40 y el otro de 50 alumnos, se les puso un examen. En el primer grupo la puntuación media fue 74 y la desviación estándar 8; en el segundo grupo la puntuación media fue 78 y la desviación estándar 7. ¿Existe diferencia en el desempeño de estos dos grupos a los niveles de significancia: *a*) 0.05 y *b*) 0.01?

SOLUCIÓN

Supóngase que los dos grupos provienen de dos poblaciones cuyas medias son μ_1 y μ_2, respectivamente. Entonces se debe decidir entre las hipótesis:

$H_0 : \mu_1 = \mu_2$, la diferencia se debe únicamente a la casualidad.

$H_1 : \mu_1 \neq \mu_2$, existe una diferencia significativa entre los dos grupos.

De acuerdo con la hipótesis H_0, ambos grupos provienen de una misma población. La media y la desviación estándar de la diferencia entre las medias están dadas por

$$\mu_{\bar{X}1-\bar{X}2} = 0 \qquad \text{y} \qquad \sigma_{\bar{X}1-\bar{X}2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{8^2}{40} + \frac{7^2}{50}} = 1.606$$

donde se han empleado las desviaciones estándar muestrales como estimación de σ_1 y σ_2. Por lo tanto,

$$z = \frac{\bar{X}_1 - \bar{X}}{\sigma_{\bar{X}1-\bar{X}2}} = \frac{74 - 78}{1.606} = -2.49$$

a) En una prueba de dos colas, los resultados son significativos al nivel 0.05 si z se encuentra fuera del intervalo de -1.96 a 1.96. Por lo tanto, se concluye que al nivel de significancia 0.05 existe una diferencia significativa en el desempeño de estos dos grupos y que el segundo grupo parece ser mejor.

b) En una prueba de dos colas, los resultados son significativos al nivel 0.01 si z se encuentra fuera del intervalo de -2.58 a 2.58. Por lo tanto, se concluye que al nivel 0.01 no hay diferencia significativa entre las clases.

Ya que los resultados son significativos al nivel 0.05 pero no al nivel 0.01, se concluye que los resultados sean *probablemente significativos* (de acuerdo con la terminología presentada al final del problema 10.5).

10.18 La estatura media de 50 estudiantes que mostraron una participación especial en las actividades deportivas de su escuela fue 68.2 pulgadas (in) con una desviación estándar de 2.5 in, en tanto que la estatura media de 50 estudiantes que no mostraron interés en los deportes fue 67.5 in con una desviación estándar de 2.8 in. Probar la hipótesis de que los estudiantes que mostraron interés en el deporte son más altos que el resto de los estudiantes.

SOLUCIÓN

Hay que decidir entre las hipótesis:

$H_0: \mu_1 = \mu_2$, no hay diferencia entre las estaturas medias.

$H_1: \mu_1 > \mu_2$, la estatura media del primer grupo es mayor que la del segundo grupo.

Bajo la hipótesis H_0,

$$\mu_{\bar{X}1-\bar{X}2} = 0 \qquad \text{y} \qquad \sigma_{\bar{X}1-\bar{X}2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \sqrt{\frac{(2.5)^2}{50} + \frac{(2.8)^2}{50}} = 0.53$$

donde para estimar σ_1 y σ_2 se han empleado las desviaciones estándar muestrales. Por lo tanto,

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}1-X2}} = \frac{68.2 - 67.5}{0.53} = 1.32$$

Usando una prueba de una cola al nivel de significancia 0.05 se puede rechazar H_0 si la puntuación z es mayor a 1.645. Por lo tanto, en este caso, a ese nivel de significancia no se puede rechazar la hipótesis nula.

Sin embargo, hay que observar que la hipótesis se puede rechazar al nivel de significancia 0.10 si se está dispuesto a correr el riesgo de tener una probabilidad de 0.10 de cometer un error (es decir, 1 posibilidad en 10).

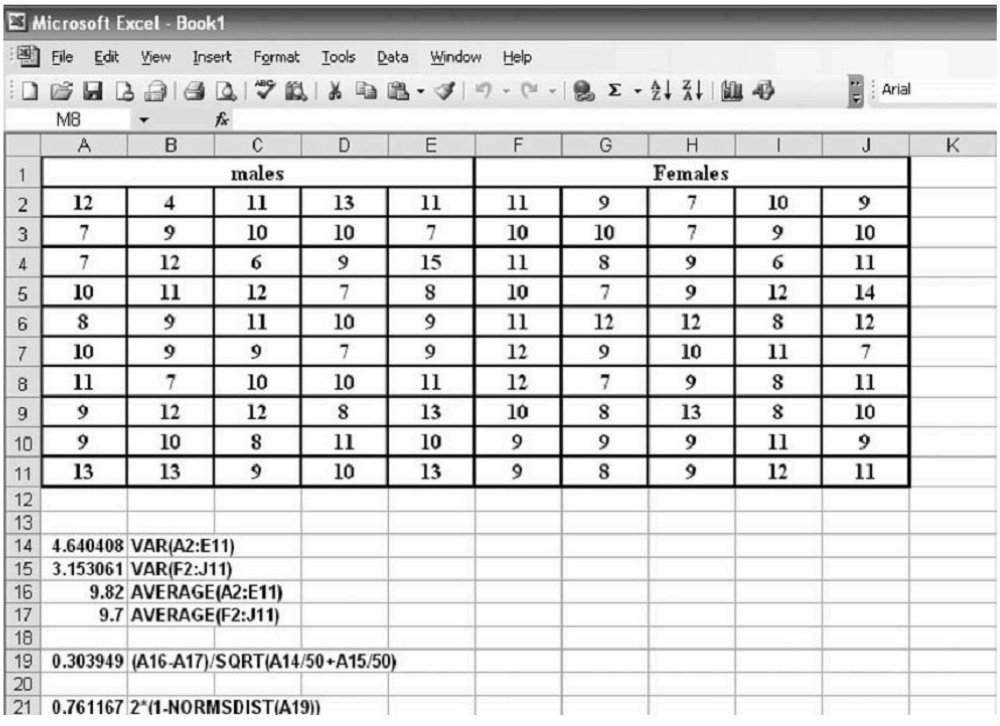
10.19 Se realiza un estudio para comparar la media, en horas por semana, que usan sus celulares varones y mujeres estudiantes universitarios. De una universidad se tomaron 50 estudiantes mujeres y 50 estudiantes varones y se registró la cantidad de horas por semana que utilizan sus celulares. Los resultados se muestran en la tabla 10.6. Se quiere probar $H_0: \mu_1 - \mu_2 = 0$ contra $H_a: \mu_1 - \mu_2 \neq 0$, basándose en estas muestras. Usar EXCEL para calcular el valor p y llegar a una decisión acerca de la hipótesis nula.

Tabla 10.6 Horas por semana que usan su celular varones y mujeres estudiantes de una universidad

Varones					Mujeres				
12	4	11	13	11	11	9	7	10	9
7	9	10	10	7	10	10	7	9	10
7	12	6	9	15	11	8	9	6	11
10	11	12	7	8	10	7	9	12	14
8	9	11	10	9	11	12	12	8	12
10	9	9	7	9	12	9	10	11	7
11	7	10	10	11	12	7	9	8	11
9	12	12	8	13	10	8	13	8	10
9	10	8	11	10	9	9	9	11	9
13	13	9	10	13	9	8	9	12	11

SOLUCIÓN

Los datos de la tabla 10.6 se ingresan en una hoja de cálculo de EXCEL como se muestra en la figura 10-12. Los datos de los varones se ingresan en las celdas A2:E11 y los datos de las mujeres en las celdas F2:J11. La varianza de los datos de los varones se calcula ingresando en la celda A14 =VAR(A2:E11). La varianza de los datos de las mujeres se calcula ingresando en la celda A15 =VAR(F2:J11). La media de los datos de los varones se calcula ingresando en la celda



	A	B	C	D	E	F	G	H	I	J	K
1	males					Females					
2	12	4	11	13	11	11	9	7	10	9	
3	7	9	10	10	7	10	10	7	9	10	
4	7	12	6	9	15	11	8	9	6	11	
5	10	11	12	7	8	10	7	9	12	14	
6	8	9	11	10	9	11	12	12	8	12	
7	10	9	9	7	9	12	9	10	11	7	
8	11	7	10	10	11	12	7	9	8	11	
9	9	12	12	8	13	10	8	13	8	10	
10	9	10	8	11	10	9	9	9	11	9	
11	13	13	9	10	13	9	8	9	12	11	
12											
13											
14	4.640408	VAR(A2:E11)									
15	3.153061	VAR(F2:J11)									
16	9.82	AVERAGE(A2:E11)									
17	9.7	AVERAGE(F2:J11)									
18											
19	0.303949	(A16-A17)/SQRT(A14/50+A15/50)									
20											
21	0.761167	2*(1-NORMSDIST(A19))									

Figura 10-12 Hoja de cálculo EXCEL para calcular el valor *p* del problema 10.19.

A16 =AVERAGE(A2:E11). La media de los datos de las mujeres se calcula ingresando en la celda A17 =AVERAGE(F2:J11). El estadístico de prueba es =(A16-A17)/SQRT(A14/50+A15/50) y se muestra en A19. Este estadístico tiene una distribución normal estándar y su valor es 0.304. La expresión =2*(1-NORMSDIST(A19) calcula el área a la derecha de 0.304 y la duplica. Con esto se obtiene que el valor $p = 0.761$.

Como este valor p no es menor que ninguno de los valores α usuales, 0.01 o bien 0.05, no se rechaza la hipótesis nula. La probabilidad de obtener muestras como la obtenida es 0.761, suponiendo que la hipótesis nula sea verdadera. Por lo tanto, no hay evidencia que sugiera que la hipótesis nula es falsa y que se deba rechazar.

10.20 Se tienen dos grupos de personas, *A* y *B*, cada uno de 100 personas que padecen una enfermedad. Al grupo *A* se le administra un suero, pero al grupo *B* (que es el grupo *control*) no; por lo demás, los dos grupos se tratan en forma idéntica. En los grupos *A* y *B* se encuentra que 75 y 65 personas, respectivamente, se recuperan de esta enfermedad. A los niveles de significancia: *a*) 0.01, *b*) 0.05 y *c*) 0.10, probar la hipótesis de que el suero ayuda a la curación de la enfermedad. Calcular el valor p y mostrar que valor $p > 0.01$, valor $p > 0.05$, pero valor $p < 0.10$.

SOLUCIÓN

Sean p_1 y p_2 las proporciones poblacionales de las personas curadas: 1) usando el suero y 2) sin usar el suero, respectivamente. Hay que decidir entre las hipótesis:

$H_0 : p_1 = p_2$, las diferencias observadas se deben a la casualidad (es decir, el suero no es eficiente).

$H_1 : p_1 > p_2$, el suero sí es eficiente.

Bajo la hipótesis H_0,

$$\mu_{P1-P2} = 0 \quad \text{y} \qquad \sigma_{P1-P2} = \sqrt{pq\left(\frac{1}{N_1}+\frac{1}{N_2}\right)} = \sqrt{(0.70)(0.30)\left(\frac{1}{100}+\frac{1}{100}\right)} = 0.0648$$

donde, como estimación de p, se ha empleado la proporción promedio de curados en las dos muestras dada por $(75 + 65)/200 = 0.70$, de donde $q = 1 - p = 0.30$. Por lo tanto,

$$z = \frac{P_1 - P_2}{\sigma_{P1-P2}} = \frac{0.750 - 0.650}{0.0648} = 1.54$$

a) Empleando una prueba de una cola al nivel de significancia 0.01, la hipótesis H_0 se rechaza únicamente si la puntuación z es mayor a 2.33. Como la puntuación z es de sólo 1.54, se concluye que a este nivel de significancia los resultados se deben a la casualidad.

b) Empleando una prueba de una cola al nivel de significancia 0.05, la hipótesis H_0 se rechaza únicamente si la puntuación z es mayor a 1.645. Por lo tanto, se concluye que a este nivel de significancia los resultados se deben a la casualidad.

c) Si se usa una prueba de una cola al nivel de significancia 0.10, H_0 se rechaza sólo si la puntuación z es mayor a 1.28. Dado que esta condición se satisface, se concluye que el suero es eficiente al nivel 0.10.

d) Empleando EXCEL, el valor p se obtiene mediante =1-NORMDIST(1.54), que es igual a 0.06178. Ésta es el área a la derecha de 1.54. Obsérvese que este valor es mayor a 0.01, 0.05, pero menor a 0.10.

Nótese que la conclusión depende de qué tanto se está dispuesto a arriesgarse a estar equivocado. Si en realidad los resultados se deben a la casualidad, pero se concluye que se deben al suero (error tipo I), se procederá a administrar el suero a una gran cantidad de personas, con el único resultado de que en realidad no sea efectivo. Éste es un riesgo que no siempre se está dispuesto a asumir.

Por otro lado, se puede concluir que el suero no ayuda, cuando en realidad sí lo hace (error tipo II). Esta conclusión es muy peligrosa, en especial porque lo que está en juego son vidas humanas.

10.21 Repetir el problema 10.20, pero considerando que cada grupo consta de 300 personas y que sanan 225 personas del grupo *A* y 195 del grupo *B*. Encontrar el valor p usando EXCEL y comentar sobre su decisión.

SOLUCIÓN

Obsérvese que la proporción de personas que sana en cada grupo es $225/300 = 0.750$ y $195/300 = 0.650$, respectivamente, que son las mismas que en el problema 10.20. De acuerdo con la hipótesis H_0

$$\mu_{P1-P2} = 0 \qquad \text{y} \qquad \sigma_{P1-P2} = \sqrt{pq\left(\frac{1}{N_1}+\frac{1}{N_2}\right)} = \sqrt{(0.70)(0.30)\left(\frac{1}{300}+\frac{1}{300}\right)} = 0.0374$$

donde $(225 + 195)/600 = 0.70$ se usa como estimación de p. Por lo tanto,

$$z = \frac{P_1 - P_2}{\sigma_{P1-P2}} = \frac{0.750 - 0.650}{0.0374} = 2.67$$

Como el valor de z es mayor que 2.33, la hipótesis nula se puede rechazar al nivel de significancia 0.01; es decir, se puede concluir que el suero es efectivo con una probabilidad de estar equivocado de sólo 0.01.

Esto muestra cómo al aumentar el tamaño de la muestra se incrementa la confiabilidad de las decisiones. Sin embargo, en muchos casos suele no ser posible aumentar el tamaño de la muestra. En esos casos se está forzado a tomar las decisiones con base en la información disponible, y por lo tanto se debe conformar con correr mayor riesgo de tomar una decisión incorrecta.

valor p =1-NORMDIST(2.67) = 0.003793. Esto es menor a 0.01.

10.22 Se realizó un sondeo en una muestra de 300 votantes del distrito *A* y 200 votantes del distrito *B*; se encontró que 56 y 48%, respectivamente, estaban a favor de determinado candidato. Al nivel de significancia 0.05, probar las hipótesis: *a*) existe diferencia entre los distritos, *b*) el candidato se prefiere en el distrito *A* y *c*) calcular el valor *p* de los incisos *a*) y *b*).

SOLUCIÓN

Sean p_1 y p_2 las proporciones de todos los votantes de los distritos *A* y *B* que están a favor de este candidato. Bajo la hipótesis $H_0 : p_1 = p_2$, se tiene

$$\mu_{P1-P2} = 0 \qquad \text{y} \qquad \sigma_{P1-P2} = \sqrt{pq\left(\frac{1}{N_1}+\frac{1}{N_2}\right)} = \sqrt{(0.528)(0.472)\left(\frac{1}{300}+\frac{1}{200}\right)} = 0.0456$$

donde se emplean los valores $[(0.56)(300) + (0.48)(200)]/500 = 0.528$ y $(1 - 0.528) = 0.472$ como estimaciones de p y q, respectivamente. Por lo tanto,

$$z = \frac{P_1 - P_2}{\sigma_{P1-P2}} = \frac{0.560 - 0.480}{0.0456} = 1.75$$

a) Si sólo se desea determinar si existe alguna diferencia entre los distritos, hay que decidir entre las hipótesis $H_0 : p_1 = p_2$ y $H_1 : p_1 \neq p_2$, lo que implica una prueba de dos colas. Usando una prueba de dos colas al nivel de significancia 0.05, H_0 se puede rechazar si z está fuera del intervalo -1.96 a 1.96. Como $z = 1.75$ se encuentra en este intervalo, a este nivel no se puede rechazar H_0; esto es, no hay diferencia significativa entre los dos distritos.

b) Si se desea determinar si el candidato es preferido en el distrito *A*, hay que decidir entre las hipótesis $H_0 : p_1 = p_2$ y $H_1 : p_1 > p_2$, lo que implica una prueba de una cola. Usando una prueba de una cola al nivel de significancia 0.05, H_0 se rechaza si z es mayor a 1.645. Dado que éste es el caso, se rechaza H_0 a este nivel de significancia y se concluye que el candidato es preferido en el distrito *A*.

c) Con la alternativa de dos colas, el valor p =2*(1-NORMDIST(1.75)) = 0.0801. A $\alpha = 0.05$ no se puede rechazar la hipótesis nula. Con la alternativa de una cola, valor p =1-NORMDIST(1.75) = 0.04006. A $\alpha = 0.05$ se puede rechazar la hipótesis nula.

PRUEBAS EMPLEANDO DISTRIBUCIONES BINOMIALES

10.23 Un profesor aplica un pequeño examen en el que hay 10 preguntas de verdadero o falso. Para probar la hipótesis de que los alumnos contestan sólo adivinando, el profesor adopta la siguiente regla de decisión:

Si hay siete o más de las respuestas correctas, el estudiante no está sólo adivinando.
Si hay menos de siete respuestas correctas, el estudiante está sólo adivinando.

Encontrar la probabilidad de rechazar la hipótesis nula cuando ésta sea correcta: *a*) empleando la distribución binomial y *b*) empleando EXCEL.

SOLUCIÓN

a) Sea p la probabilidad de que una pregunta se responda correctamente. La probabilidad de tener X de 10 preguntas correctas es $\binom{10}{X}p^X q^{10-X}$, donde $q = 1 - p$. Entonces bajo la hipótesis $p = 0.5$ (es decir, el estudiante está sólo atinando),

$$\Pr\{7 \text{ o más correctas}\} = \Pr\{7 \text{ correctas}\} + \Pr\{8 \text{ correctas}\} + \Pr\{9 \text{ correctas}\} + \Pr\{10 \text{ correctas}\}$$

$$= \binom{10}{7}\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right)^3 + \binom{10}{8}\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^2 + \binom{10}{9}\left(\frac{1}{2}\right)^9\left(\frac{1}{2}\right) + \binom{10}{10}\left(\frac{1}{2}\right)^{10} = 0.1719$$

Por lo tanto, la probabilidad de concluir que el estudiante no está sólo adivinando cuando en realidad sí lo esté haciendo es 0.1719. Obsérvese que ésta es la probabilidad de un error tipo I.

b) Los números 7, 8, 9 y 10 se ingresan en A1:A4 de la hoja de cálculo de EXCEL. Después se ingresa =BINOMDIST(A1, 10,0.5,0). A continuación se hace clic y se arrastra desde B1 hasta B4. En B5 se ingresa =SUM(B1:B4). La respuesta aparece en B5.

A	B
7	0.117188
8	0.043945
9	0.009766
10	0.000977
	0.171875

10.24 En el problema 10.23, encontrar la probabilidad de aceptar la hipótesis $p = 0.5$ cuando en realidad $p = 0.7$. Encontrar la respuesta: *a*) usando la fórmula de probabilidad binomial y *b*) usando EXCEL.

SOLUCIÓN

a) Bajo la hipótesis $p = 0.7$,

$$\Pr\{\text{menos de 7 correctas}\} = 1 - \Pr\{7 \text{ o más correctas}\}$$

$$= 1 - \left[\binom{10}{7}(0.7)^7(0.3)^3 + \binom{10}{8}(0.7)^8(0.3)^2 + \binom{10}{9}(0.7)^9(0.3) + \binom{10}{10}(0.3)^{10}\right]$$

$$= 0.3504$$

b) La solución usando EXCEL es:

Pr{menos de 7 correctas cuando $p = 0.7$} está dada por =BINOMDIST (6,10,0.7,1) que es igual a 0.350389. El 1 en la función BINOMDIST indica que la probabilidad, correspondiente a $n = 10$ y $p = 0.7$, desde 0 hasta 6 está acumulada.

10.25 En el problema 10.23, encontrar la probabilidad de aceptar la hipótesis $p = 0.5$ cuando en realidad: *a*) $p = 0.6$, *b*) $p = 0.8$, *c*) $p = 0.9$, *d*) $p = 0.4$, *e*) $p = 0.3$, *f*) $p = 0.2$ y *g*) $p = 0.1$.

SOLUCIÓN

a) Si $p = 0.6$,

$$\text{Probabilidad buscada} = 1 - [\Pr\{7 \text{ correctas}\} + \Pr\{8 \text{ correctas}\} + \Pr\{9 \text{ correctas}\} + \Pr\{10 \text{ correctas}\}]$$

$$= 1 - \left[\binom{10}{7}(0.6)^7(0.4)^3 + \binom{10}{8}(0.6)^8(0.4)^2 + \binom{10}{9}(0.6)^9(0.4) + \binom{10}{10}(0.6)^{10}\right] = 0.618$$

Los resultados de los incisos *b*) a *g*) se encuentran de manera similar y se presentan en la tabla 10.7, junto con los correspondientes valores desde $p = 0.5$ hasta $p = 0.7$. Obsérvese que en la tabla 10.7 la probabilidad se denota por β (probabilidad de cometer un error tipo II); la entrada β correspondiente a $p = 0.5$ está dada por $\beta = 1 - 0.1719 = 0.828$ (de acuerdo con el problema 10.23) y la entrada β correspondiente a $p = 0.7$ se tomó del problema 10.24.

Tabla 10.7

p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
β	1.000	0.999	0.989	0.945	0.828	0.618	0.350	0.121	0.013

10.26 Usar el problema 10.25 para construir una gráfica de β contra p.

SOLUCIÓN

La gráfica buscada se muestra en la figura 10-3.

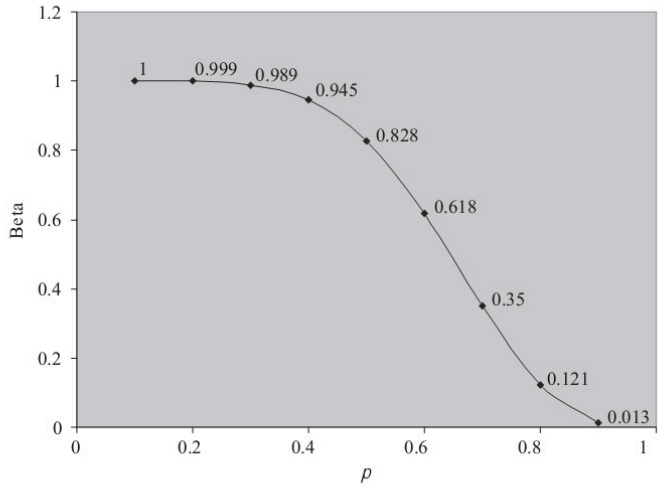


Figura 10-13 Gráfica para los errores tipo II en el problema 10.25.

10.27 La hipótesis nula es que un dado no está cargado y la hipótesis alternativa es que el dado sí está cargado, de manera que la cara seis aparece con más frecuencia de la que debería. Esta hipótesis se prueba lanzando el dado 18 veces y observando cuántas veces cae seis. Encontrar el valor p si la cara seis se presenta 7 veces en 18 lanzamientos del dado.

SOLUCIÓN

En la hoja de cálculo de EXCEL se ingresan en A1:A19 los números del 0 al 18. En B1 se ingresa =BINOMDIST(A1,18, 0.16666,0), se hace clic y se arrastra desde B1 hasta B19 para obtener cada una de las probabilidades binomiales, en C1 se ingresa =BINOMDIST (A1,18,0.16666,1), se hace clic y se arrastra desde C1 hasta C19 con lo que se obtiene la probabilidad binomial acumulada.

A	B	C
0	0.037566	0.037566446
1	0.135233	0.17279916
2	0.229885	0.402683738
3	0.245198	0.647882186
4	0.18389	0.831772194
5	0.102973	0.934745656
6	0.04462	0.979365347
7	0.015297	0.994662793
8	0.004207	0.998869389
9	0.000935	0.999804143
10	0.000168	0.999972391
11	2.45E-05	0.999996862
12	2.85E-06	0.999999717
13	2.64E-07	0.99999998
14	1.88E-08	0.999999999
15	1E-09	1
16	3.76E-11	1
17	8.86E-13	1
18	9.84E-15	1

El valor p es $p\{x \geq 7\} = 1 - P\{X \leq 6\} = 1 - 0.979 = 0.021$. El resultado $X = 6$ es significativo a $\alpha = 0.05$, pero no a $\alpha = 0.01$.

10.28 Para probar que 40% de las personas que pagan impuestos emplean algún software para el cálculo de los mismos contra la hipótesis alternativa de que el porcentaje es mayor a 40%, se seleccionan en forma aleatoria 300 personas que pagan impuestos y se les pregunta si emplean algún software. Si 131 de las 300 emplea algún software, encontrar el valor p correspondiente a esta observación.

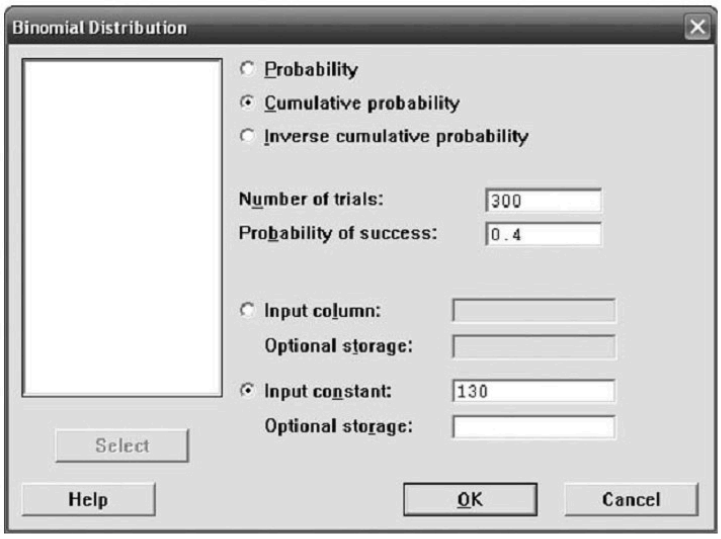


Figura 10-14 Cuadro de diálogo de la distribución binomial para calcular 130 o menos de 300 usuarios de software, dado que 40% de los que pagan impuestos usan algún software.

SOLUCIÓN

La hipótesis nula es $H_0: p = 0.4$ y la hipótesis alternativa $H_a: p > 0.4$. El valor de X observado es 131, donde X es la cantidad de los que usan algún software. El valor $p = P\{X \geq 131 \text{ dado que } p = 0.4\}$. El valor $p = 1 - P\{X \leq 130 \text{ dado que } p = 0.4\}$. Empleando MINITAB, con la secuencia **"Calc ⇒ Probability Distribution ⇒ Binomial"** se abre el cuadro de diálogo que se muestra en la figura 10-14.

Con el cuadro de diálogo de la figura 10-14 se obtiene el resultado siguiente.

Función de distribución acumulada

```
Binomial with n=300 and p=0.4

x       P(X<=x)
130     0.891693
```

El valor p es $1 - P\{X \leq 130 \text{ dado que } p = 0.4\} = 1 - 0.8971 = 0.1083$. El resultado $X = 131$ no es significativo a 0.01, 0.05 ni 0.10.

PROBLEMAS SUPLEMENTARIOS

PRUEBAS PARA MEDIAS Y PARA PROPORCIONES EMPLEANDO DISTRIBUCIONES NORMALES

10.29 Una urna contiene sólo canicas azules y rojas. Para probar la hipótesis nula de que las canicas de ambos colores se encuentran en la misma proporción, se toma una muestra, con reposición, de 64 canicas; se anotan los colores de las canicas que se van extrayendo y se adopta la siguiente regla de decisión:

La hipótesis nula se acepta si $28 \leq X \leq 36$, donde X es la cantidad de canicas rojas en la muestra de tamaño 64.

La hipótesis nula se rechaza si $X \leq 27$ o si $X \geq 37$.

a) Encontrar la probabilidad de rechazar la hipótesis nula si es correcta.

b) Graficar la regla de decisión y el resultado que se obtenga en el inciso *a*).

10.30 *a*) ¿Qué regla de decisión se adopta en el problema 10.29 si lo que se busca es que la probabilidad de rechazar la hipótesis nula siendo en realidad correcta no sea mayor a 0.01 (es decir, si se quiere que el nivel de significancia sea 0.01)?

b) ¿A qué nivel de confianza se puede aceptar la hipótesis nula?

c) ¿Cuál es la regla de decisión si se emplea como nivel de significancia 0.05?

10.31 Supóngase que en el problema 10.29 se desea probar la hipótesis de que la proporción de canicas rojas es mayor que la de canicas azules.

a) ¿Cuál es entonces la hipótesis nula y cuál la hipótesis alternativa?

b) ¿Se debe usar una prueba de una cola o de dos colas? ¿Por qué?

c) ¿Cuál debe ser la regla de decisión si el nivel de significancia es 0.05?

d) ¿Cuál es la regla de decisión si el nivel de significancia es 0.01?

10.32 Se lanzan 100 veces un par de dados y en 23 de las veces aparece un 7. Al nivel de significancia 0.05, probar la hipótesis de que los dados no están cargados empleando: *a*) una prueba de dos colas y *b*) una prueba de una cola. Analizar las razones, si es que las hay, para preferir una de estas dos pruebas.

10.33 Repetir el problema 10.32 empleando como nivel de significancia 0.01.

10.34 Un fabricante asegura que por lo menos el 95% de los equipos que vende a una fábrica satisfacen las especificaciones. En una muestra de 200 equipos examinados, 18 no cumplen con las especificaciones. Probar la afirmación del fabricante a los niveles de significancia: *a*) 0.01 y *b*) 0.05.

10.35 Se afirma que los compradores por Internet gastan en promedio $335 por año. Se desea probar que esta cantidad no es la correcta empleando $\alpha = 0.075$. Se hace un estudio en el que intervienen 300 compradores por Internet y se encuentra que la media muestral es $354 y la desviación estándar es $125. Encontrar el valor del estadístico de prueba, los valores críticos y efectuar la conclusión.

10.36 Por experiencia se sabe que la resistencia a la ruptura de determinada marca de hilo es 9.72 onzas (oz) y su desviación estándar es 1.40 oz. En una muestra reciente de 36 piezas de este hilo se encuentra que la resistencia media a la ruptura es 8.93 oz. Probar la hipótesis nula $H_0 : \mu = 9.72$ contra la hipótesis alternativa $H_0 : \mu < 9.72$ y dar el valor del estadístico de prueba y el valor crítico que corresponde a: *a*) $\alpha = 0.10$ y *b*) $\alpha = 0.025$. ¿Es este resultado significativo a $\alpha = 0.10$? ¿Es este resultado significativo a $\alpha = 0.025$?

10.37 Se realiza un estudio para probar la hipótesis nula de que la cantidad media de correos electrónicos enviados semanalmente por los empleados en una ciudad grande es 25.5 contra la hipótesis alternativa de que esta cantidad es mayor a 25.5. Se entrevista a 200 empleados de toda la ciudad y se encuentra que $\bar{x} = 30.1$ y $s = 10.5$. Dar el valor del estadístico de prueba y el valor crítico para $\alpha = 0.03$, y efectuar la conclusión.

10.38 Para una n grande ($n > 30$) y una desviación estándar conocida se usa la distribución normal estándar para realizar una prueba acerca de la media de la población de la que se toma la muestra. A la hipótesis alternativa $H_a : \mu < \mu_0$ se le llama *alternativa de la cola inferior* y a la hipótesis alternativa $H_a : \mu > \mu_0$ se le llama *alternativa de la cola superior*. Para una alternativa de la cola superior, dar la expresión de EXCEL para el valor crítico si $\alpha = 0.1$, $\alpha = 0.01$ y $\alpha = 0.001$.

VALORES *p* EN PRUEBAS DE HIPÓTESIS

10.39 Para probar que una moneda está balanceada se lanza 15 veces y se obtienen 12 caras. Dar el valor p correspondiente a este resultado. Para hallar el valor p emplear BINOMDIST de EXCEL.

10.40 Dar el valor p correspondiente al resultado del problema 10.35.

10.41 Dar el valor p correspondiente al resultado del problema 10.36.

10.42 Dar el valor p correspondiente al resultado del problema 10.37.

GRÁFICAS DE CONTROL DE CALIDAD

10.43 Cierto tipo de hilo producido por un fabricante ha tenido una resistencia a la ruptura de 8.64 oz y una desviación estándar de 1.28 oz. Para determinar si este producto satisface los estándares, cada tres horas se toma una muestra de 16 piezas y se determina la media de su resistencia al rompimiento. En una gráfica de control de calidad, registrar los límites de control de: *a*) 99.73% (o 3σ), *b*) 99% y *c*) 95%, y explicar sus aplicaciones.

10.44 En promedio, cerca del 3% de los pernos que produce una empresa están defectuosos. Para mantener esta calidad cada cuatro horas se toma una muestra de 200 pernos y se examina. Determinar los límites de control de: *a*) 99% y *b*) 95% para la cantidad de pernos defectuosos en cada muestra. Obsérvese que en este caso sólo se necesitan los *límites superiores de control*.

PRUEBAS PARA DIFERENCIAS DE MEDIAS Y PROPORCIONES

10.45 En un estudio se compara la vida media, en horas, de dos tipos de focos. Los resultados del estudio se muestran en la tabla 10.8.

Tabla 10.8

	Foco ecológico	Foco tradicional
n	75	75
Media	1 250	1 305
Desv. est.	55	65

Probar $H_0 : \mu_1 - \mu_2 = 0$ contra $H_a : \mu_1 - \mu_2 \neq 0$ con $\alpha = 0.05$. Dar el valor de la prueba estadística y calcular el valor de p y comparar el valor p con $\alpha = 0.05$. Proporcionar su conclusión.

10.46 En un estudio se comparan las calificaciones de 50 estudiantes universitarios que tienen televisión en su dormitorio con las de 50 estudiantes universitarios que no tienen televisión en su dormitorio. Los resultados se muestran en la tabla 10.9. La hipótesis alternativa es que la media de las calificaciones de los universitarios que no tienen televisión en su dormitorio es mayor a la de los que sí la tienen. Dar el valor del estadístico de prueba suponiendo que no haya diferencia entre las calificaciones. Dar el valor p y las conclusiones para $\alpha = 0.05$ y para $\alpha = 0.10$.

Tabla 10.9

	Televisión en el dormitorio	Sin televisión
n	50	50
Media	2.58	2.77
Desv. est.	0.55	0.65

10.47 En un examen de ortografía en una escuela primaria, la calificación promedio de 32 niños fue de 72 puntos y su desviación estándar de 8 puntos, y la calificación promedio de 36 niñas fue de 75 puntos y su desviación estándar de 6 puntos. La hipótesis alternativa es que las niñas son mejores en ortografía que los niños. Dar el valor del estadístico de prueba suponiendo que entre niños y niñas no hay diferencia en la calificación de ortografía. Dar el valor p y la conclusión para $\alpha = 0.05$ y para $\alpha = 0.10$.

10.48 Para probar los efectos de un nuevo fertilizante sobre la producción de trigo se dividió una parcela en 60 cuadrados de la misma área, todos de idéntica calidad en términos de suelo, exposición a la luz, etc. En 30 de los cuadrados se empleó el nuevo fertilizante y el fertilizante viejo se usó en el resto de los cuadrados. La cantidad media de bushels (bu) de trigo, usando el nuevo fertilizante, cosechado por cuadrado, fue de 18.2 bu y su desviación estándar de 0.63 bu. La media y la desviación estándar correspondientes en el caso en que se usó el fertilizante viejo fueron 17.8 y 0.54 bu, respectivamente. Empleando como niveles de significancia: *a*) 0.05 y *b*) 0.01, probar la hipótesis de que el nuevo fertilizante es mejor que el viejo.

10.49 En muestras aleatorias de 200 remaches elaborados con la máquina A y 100 remaches elaborados con la máquina B se encontraron 19 y 5 remaches defectuosos, respectivamente.

a) Dar el estadístico de prueba, el valor p, y su conclusión a $\alpha = 0.05$ para probar que las dos máquinas tienen diferente calidad de desempeño.

b) Dar el estadístico de prueba, el valor p y la conclusión a $\alpha = 0.05$ para probar que la máquina B es mejor que la máquina A.

10.50 Dos urnas, A y B, contienen la misma cantidad de canicas, pero no se sabe cuál es la proporción de canicas rojas y canicas blancas en cada una de ellas. De cada una se toma una muestra, con reposición, de 50 canicas. En las 50 canicas de la urna A hay 32 rojas y en las 50 canicas de la urna B hay 23 rojas.

a) Empleando $\alpha = 0.05$, probar la hipótesis de que la proporción de canicas rojas es la misma en las dos urnas, contra la hipótesis de que es diferente; dar el estadístico de prueba calculado, el valor p calculado y la conclusión.

b) Empleando $\alpha = 0.05$, probar la hipótesis de que la urna A tiene una proporción mayor de canicas rojas que la urna B; dar el estadístico de prueba calculado, el valor p calculado y la conclusión.

10.51 Para determinar si una moneda está cargada, de manera que al lanzarla sea más probable que aparezca cara que cruz, se lanza 15 veces. Sea $X =$ cantidad de caras en los 15 lanzamientos. Se declarará que la moneda está cargada a favor de cara si $X \geq 11$. Usar EXCEL para hallar α.

10.52 Se lanza una moneda 20 veces para determinar si está cargada. Se declarará cargada si $X = 0, 1, 2, 18, 19, 20$, donde $X =$ cantidad de cruces obtenidas. Usar EXCEL para hallar α.

10.53 Se lanza una moneda 15 veces para determinar si está cargada, de manera que al lanzarla sea más probable que aparezca cara que cruz. Sea $X =$ cantidad de caras en los 15 lanzamientos. Se declarará cargada a favor de cara si $X \geq 11$. Usar EXCEL y encontrar β si $p = 0.6$.

10.54 Se lanza una moneda 20 veces para determinar si está cargada. Se declarará cargada si $X = 0, 1, 2, 18, 19, 20$, donde $X =$ cantidad de cruces obtenidas. Usar EXCEL para hallar β si $p = 0.9$.

10.55 Se lanza una moneda 15 veces para determinar si está cargada, de manera que al lanzarla sea más probable que aparezca cara que cruz. Sea $X =$ cantidad de caras en los 15 lanzamientos. Se declarará cargada a favor de cara si $X \geq 11$. Encontrar el valor p correspondiente al resultado $X = 10$. Comparar el valor p con el valor de α en este problema.

10.56 Se lanza una moneda 20 veces para determinar si está cargada. Se declarará cargada si $X = 0, 1, 2, 3, 4, 16, 17, 18, 19$ y 20, donde $X =$ cantidad de cruces obtenidas. Encontrar el valor p correspondiente al resultado $X = 17$. Comparar el valor p con el valor de α en este problema.

10.57 En una línea de producción se fabrican teléfonos celulares. Tres por ciento de defectuosos se considera aceptable. De la producción diaria se selecciona una muestra de 50. Si en la muestra se encuentran más de tres defectuosos, se considera que el porcentaje de defectuosos se ha excedido del 3% y la línea de producción se detiene hasta que se satisfaga el 3%. Emplear EXCEL para determinar α.

10.58 En el problema 10.57 encontrar la probabilidad de que 4% de defectuosos no haga que se detenga la línea de producción.

10.59 Para determinar si un dado está balanceado se lanza 20 veces. Se declarará que no está balanceado porque el 6 aparece más de $1/6$ de las veces si en 20 lanzamientos se obtienen más de 5 seises. Hallar el valor de α. Si se lanza el dado 20 veces y se obtienen 6 seises, hallar el valor p correspondiente a este resultado.

TEORÍA DE LAS MUESTRAS PEQUEÑAS

11

En los capítulos anteriores con frecuencia se utilizó el hecho de que si el tamaño de las muestras es grande, $N > 30$, lo que se conoce como *muestras grandes,* las distribuciones muestrales de muchos de los estadísticos son aproximadamente normales; esta aproximación mejora a medida que aumenta N. Si el tamaño de las muestras es $N < 30$, lo que se conoce como *muestras pequeñas*, esta aproximación no es buena y empeora a medida que N disminuye, de manera que es necesario hacer algunas modificaciones.

Al estudio de las distribuciones muestrales de los estadísticos, cuando las muestras son pequeñas, se le llama *teoría de las muestras pequeñas.* Sin embargo, un nombre más adecuado sería *teoría del muestreo exacto,* ya que los resultados obtenidos son válidos tanto para muestras grandes como para muestras pequeñas. En este capítulo se estudian tres distribuciones importantes: la distribución t de Student, la distribución ji cuadrada y la distribución F.

DISTRIBUCIÓN t DE STUDENT

Sea el estadístico

$$t = \frac{\bar{X} - \mu}{s}\sqrt{N-1} = \frac{\bar{X} - \mu}{\hat{s}/\sqrt{N}} \tag{1}$$

que es análogo al estadístico z dado por

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}.$$

Si se consideran muestras de tamaño N extraídas de una población normal (o aproximadamente normal) cuya media es μ y si para cada muestra se calcula t, usando la media muestral $\bar{X}$ y la desviación estándar muestral s o $\hat{s}$, se obtiene la distribución muestral de t. Esta distribución (ver figura 11-1) está dada por

$$Y = \frac{Y_0}{\left(1 + \dfrac{t^2}{N-1}\right)^{N/2}} = \frac{Y_0}{\left(1 + \dfrac{t^2}{\nu}\right)^{(\nu+1)/2}} \tag{2}$$

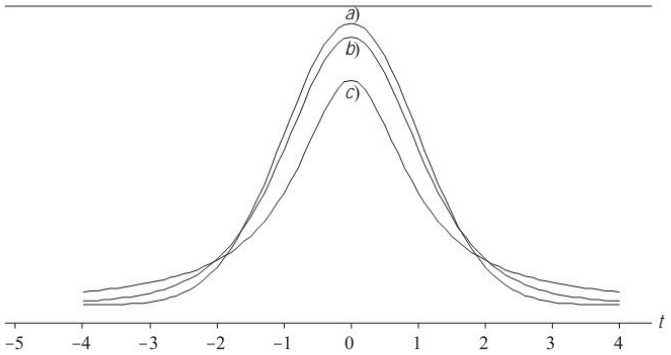


Figura 11-1 *a*) **Curva normal estándar,** *b*) *t* **de Student para** $\nu = 5$, *c*) *t* **de Student para** $\nu = 1$.

donde Y_0 es una constante que depende de N, tal que el área total bajo la curva sea 1, y donde a la constante $\nu = (N - 1)$ se le conoce como el *número de grados de libertad* (ν es la letra griega nu).

A la distribución (2) se le llama *distribución t de Student* en honor a su descubridor, W. S. Gossett, quien en la primera mitad del siglo xx publicó sus trabajos bajo el seudónimo "Student".

Si los valores de ν o de N son grandes ($N \geq 30$), la curva (2) se aproxima a la curva normal estándar

$$Y = \frac{1}{\sqrt{2\pi}} e^{-(1/2)t^2}$$

como se muestra en la figura 11-1.

INTERVALOS DE CONFIANZA

Como se hizo en el capítulo 9 con las distribuciones normales, se pueden definir intervalos de confianza de 95%, 99% u otros intervalos usando la tabla de la distribución t que aparece en el apéndice III. De esta manera puede estimarse la media poblacional μ dentro de determinados límites de confianza.

Por ejemplo, si $-t_{.975}$ y $t_{.975}$ son los valores de t para los cuales 2.5% del área se encuentra repartida en cada una de las colas de la distribución t, entonces el intervalo de confianza para t de 95% es

$$-t_{.975} < \frac{\bar{X} - \mu}{s}\sqrt{N-1} < t_{.975} \tag{3}$$

a partir de lo cual se puede estimar que μ se encuentra en el intervalo

$$\bar{X} - t_{.975}\frac{s}{\sqrt{N-1}} < \mu < \bar{X} + t_{.975}\frac{s}{\sqrt{N-1}} \tag{4}$$

con una confianza de 95% (es decir, con una probabilidad de 0.95). Obsérvese que $t_{.975}$ representa el valor del percentil 97.5, y que $t_{.025} = -t_{.975}$ representa el valor del percentil 2.5.

En general, los límites de confianza para la media poblacional se representan mediante

$$\bar{X} \pm t_c \frac{s}{\sqrt{N-1}} \tag{5}$$

donde los valores $\pm t_c$, llamados *valores críticos* o *coeficientes de confianza*, dependen del nivel de confianza deseado y del tamaño de la muestra. Estos valores se leen en el apéndice III.

Se supone que la muestra se toma de una población normal. Esta suposición se puede verificar empleando la prueba para normalidad de Kolmogorov-Smirnov.

Comparando la ecuación (5) con los límites de confianza ($\bar{X} \pm z_c\sigma/\sqrt{N}$) dados en el capítulo 9, se ve que cuando se tienen muestras pequeñas z_c (que se obtienen de la distribución normal) se sustituye por t_c (que se obtiene de la distribución *t*) y que σ se sustituye por $\sqrt{N/(N-1)}s = \hat{s}$, que es la estimación de σ. A medida que *N* aumenta, ambos métodos tienden a coincidir.

PRUEBAS DE HIPÓTESIS Y DE SIGNIFICANCIA

Las pruebas de hipótesis y de significancia, o reglas de decisión (vistas en el capítulo 10), pueden extenderse fácilmente a problemas con muestras pequeñas; la única diferencia es que la *puntuación z*, o *estadístico z*, se sustituye por la *puntuación t* o *estadístico t* apropiado.

1. **Media.** Para probar la hipótesis H_0 de que una población normal tiene una media μ, se usa la puntuación *t* (o estadístico *t*)

$$t = \frac{\bar{X} - \mu}{s}\sqrt{N-1} = \frac{\bar{X} - \mu}{\hat{s}}\sqrt{N} \tag{6}$$

donde $\bar{X}$ es la media de una muestra de tamaño *N*. Esto es análogo a usar la puntuación *z*

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}}$$

para una *N* grande, salvo que se usa $\hat{s} = \sqrt{N/(N-1)}s$ en lugar de σ. La diferencia es que mientras *z* está distribuida normalmente, *t* sigue una distribución de Student. A medida que *N* aumenta, estas distribuciones tienden a coincidir.

2. **Diferencias entre medias.** Supóngase que de poblaciones normales cuya desviaciones estándar son iguales ($\sigma_1 = \sigma_2$) se toman dos muestras aleatorias de tamaños N_1 y N_2. Supóngase, además, que las medias de estas dos muestras son $\bar{X}_1$ y $\bar{X}_2$ y que sus desviaciones estándar son s_1 y s_2, respectivamente. Para probar la hipótesis H_0 de que las muestras provienen de una misma población (es decir que $\mu_1 = \mu_2$ y también $\sigma_1 = \sigma_2$) se usa la puntuación *t* dada por

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{1/N_1 + 1/N_2}} \qquad \text{donde} \qquad \sigma = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}} \tag{7}$$

Esta distribución *t* tiene una distribución de Student con $\nu = N_1 + N_2 - 2$ grados de libertad. El uso de la ecuación (7) se hace plausible al hacer $\sigma_1 = \sigma_2 = \sigma$ en la puntuación *z* de la ecuación (2) del capítulo 10 y después usar como estimación de σ^2 la media ponderada

$$\frac{(N_1 - 1)\hat{s}_1^2 + (N_2 - 1)\hat{s}_2^2}{(N_1 - 1) + (N_2 - 1)} = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}$$

donde $\hat{s}_1^2$ y $\hat{s}_2^2$ son estimadores insesgados de σ_1^2 y σ_2^2.

DISTRIBUCIÓN JI CUADRADA

Sea el estadístico

$$\chi^2 = \frac{Ns^2}{\sigma^2} = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_N - \bar{X})^2}{\sigma^2} \tag{8}$$

donde χ es la letra griega *ji* y χ^2 se lee "ji cuadrada".

Si se consideran muestras de tamaño *N* obtenidas de una población normal cuya desviación estándar es σ, y si para cada muestra se calcula χ^2, se obtiene una distribución muestral de χ^2. Esta distribución, llamada *distribución ji cuadrada*, está dada por

$$Y = Y_0(\chi^2)^{(1/2)(\nu-2)}\, e^{-(1/2)\chi^2} = Y_0\chi^{\nu-2}\, e^{-(1/2)\chi^2} \tag{9}$$

donde $\nu = N - 1$ es el *número de grados de libertad* y Y_0 es una constante que depende de ν, de manera que el área bajo la curva sea 1. En la figura 11-2 se presentan distribuciones ji cuadrada correspondientes a diversos valores de ν. El valor máximo de Y se obtiene cuando $\chi^2 = \nu - 2$ para $\nu \geq 2$.

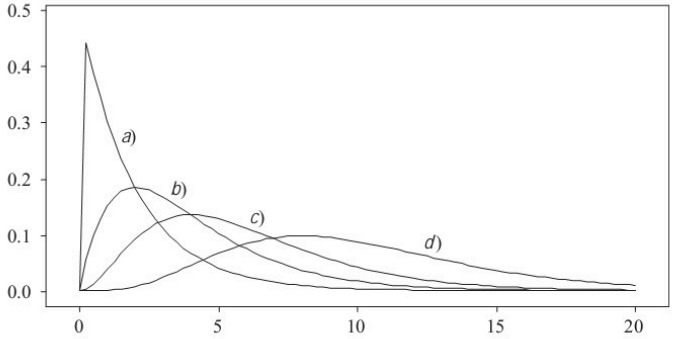


Figura 11-2 Distribuciones ji cuadrada correspondientes a: *a*) 2, *b*) 4, *c*) 6 y *d*) 10 grados de libertad.

INTERVALOS DE CONFIANZA PARA σ

Como se hizo con la distribución normal y con la distribución t, pueden definirse límites de confianza de 95%, 99%, u otros límites empleando la tabla de distribución χ^2 que se presenta en el apéndice IV. De esta manera puede estimarse la desviación estándar poblacional σ en términos de la desviación estándar muestral dentro de determinados límites de confianza.

Por ejemplo, si $\chi^2_{.025}$ y $X^2_{.975}$ son los valores de χ^2 (llamados *valores críticos*), tales que 2.5% del área se encuentra repartida en ambas colas de la distribución, entonces el intervalo de confianza de 95% es

$$\chi^2_{.025} < \frac{Ns^2}{\sigma^2} < \chi^2_{.975} \tag{10}$$

de donde se ve que puede estimarse que σ se encuentra en el intervalo

$$\frac{s\sqrt{N}}{\chi_{.975}} < \sigma < \frac{s\sqrt{N}}{\chi_{.025}} \tag{11}$$

con 95% de confianza. De manera similar se pueden encontrar otros intervalos de confianza. Los valores $\chi_{.025}$ y $\chi_{.975}$ representan, respectivamente, los percentiles 2.5 y 97.5.

En el apéndice IV se encuentran valores percentiles correspondientes a diversos grados de libertad ν. Si se tienen valores grandes de ν ($\nu \geq 30$), se puede usar el hecho de que $(\sqrt{2\chi^2} - \sqrt{2\nu - 1})$ se aproxima mucho a una distribución normal con media 0 y desviación estándar 1; por lo tanto, las tablas para la distribución normal pueden emplearse cuando $\nu \geq 30$. Si χ^2_p y z_p son los percentiles p de la distribución ji cuadrada y de la distribución normal, respectivamente, se tiene

$$\chi^2_p = \tfrac{1}{2}(z_p + \sqrt{2\nu - 1})^2 \tag{12}$$

En este caso hay una gran coincidencia con los resultados obtenidos en los capítulos 8 y 9.

Para más aplicaciones de la distribución ji cuadrada, ver el capítulo 12.

GRADOS DE LIBERTAD

Para calcular un estadístico, por ejemplo (*1*) y (*8*), es necesario emplear observaciones obtenidas de una muestra y también ciertos parámetros poblacionales. Si estos parámetros no se conocen, es necesario estimarlos a partir de la muestra.

El *número de grados de libertad* de un estadístico, que por lo general se denota ν, se define como la cantidad N de observaciones en la muestra (es decir, el tamaño de la muestra) menos la cantidad k de parámetros poblacionales que tengan que estimarse a partir de las observaciones muestrales. En símbolos, $\nu = N - k$.

En el caso del estadístico (*1*), la cantidad de observaciones independientes en la muestra es N, y a partir de ellas se calculan $\bar{X}$ y s. Sin embargo, como se necesita estimar μ, $k = 1$ y por lo tanto $\nu = N - 1$.

En el caso del estadístico (*8*), la cantidad de observaciones independientes en la muestra es N, a partir de las cuales se calcula s. Sin embargo, como se tiene que estimar σ, $k = 1$ y por lo tanto $\nu = N - 1$.

LA DISTRIBUCIÓN *F*

Según se ha visto, en algunas aplicaciones es importante conocer la distribución muestral de la diferencia entre las medias $(\bar{X}_1 - \bar{X}_2)$ de dos muestras. De igual manera, algunas veces se necesita la distribución muestral de la diferencia entre varianzas $(S_1^2 - S_2^2)$. Sin embargo, resulta que esta distribución es bastante complicada. Debido a ello, se considera el estadístico S_1^2/S_2^2, ya que un cociente grande o pequeño indica una gran diferencia, en tanto que un cociente cercano a 1 indica una diferencia pequeña. En este caso se puede encontrar una distribución muestral a la que se le conoce como *distribución F* en honor a R. A. Fischer.

Más precisamente, supóngase que se tienen dos muestras, 1 y 2, de tamaños N_1 y N_2, respectivamente, obtenidas de dos poblaciones normales (o casi normales) cuyas varianzas son σ_1^2 y σ_2^2. Sea el estadístico

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2} = \frac{N_1 S_1^2/(N_1 - 1)\sigma_1^2}{N_2 S_2^2/(N_2 - 1)\sigma_2^2} \tag{13}$$

donde

$$\hat{S}_1^2 = \frac{N_1 S_1^2}{N_1 - 1} \qquad \hat{S}_2^2 = \frac{N_2 S_2^2}{N_2 - 1}. \tag{14}$$

Entonces a la distribución muestral de F se le llama distribución F de Fisher, o simplemente distribución F, con $\nu_1 = N_1 - 1$ y $\nu_2 = N_2 - 1$ grados de libertad. Esta distribución está dada por

$$Y = \frac{CF^{(\nu_1/2)-1}}{(\nu_1 F + \nu_2)^{(\nu_1+\nu_2)/2}} \tag{15}$$

donde C es una constante que depende de ν_1 y ν_2, de manera que el área total bajo la curva sea 1. Esta curva tiene una forma similar a la de las curvas que se muestran en la figura 11-3, aunque esta forma puede variar de manera notable de acuerdo con los valores de ν_1 y ν_2.

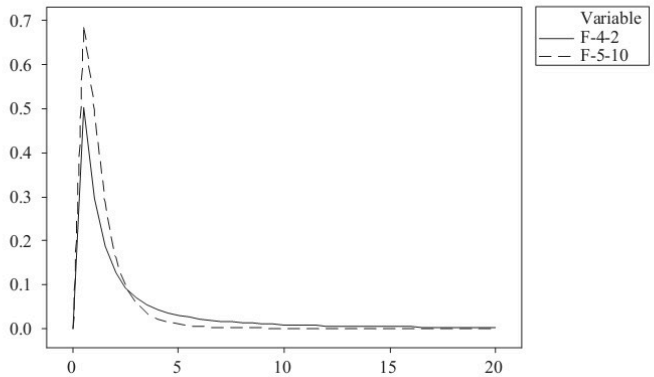


Figura 11-3 La línea continua representa la distribución *F* con 4 y 2 grados de libertad, y la línea punteada representa la distribución *F* con 5 y 10 grados de libertad.

En los apéndices V y VI se dan los valores percentiles de F para los cuales las áreas en la cola derecha son 0.05 y 0.01, respectivamente, que se denotan $F_{.95}$ y $F_{.99}$. Estos valores que representan los niveles de significancia del 5% y del 1% se usan para determinar si la varianza S_1^2 es significativamente mayor que la varianza S_2^2. En la práctica, como muestra 1 se considera la muestra que tenga la mayor varianza.

El software para estadística permite encontrar las áreas bajo la distribución t de Student, la distribución ji cuadrada y la distribución F. Este software también permite trazar las distintas distribuciones. Esto se ilustrará en la sección de problemas resueltos de este capítulo.

PROBLEMAS RESUELTOS

DISTRIBUCIÓN t DE STUDENT

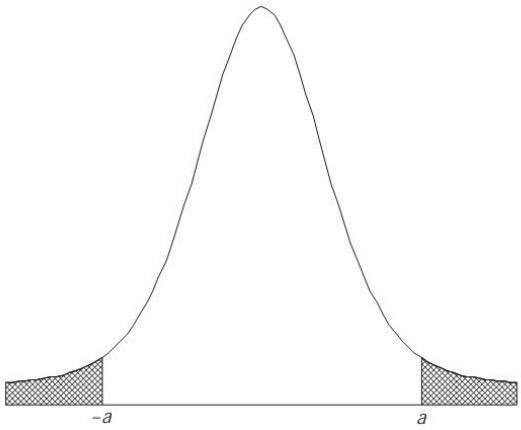


Figura 11-4 Distribución t de Student para 9 grados de libertad.

11.1 En la figura 11-4 se muestra la gráfica de la distribución t de Student para nueve grados de libertad. Utilizar el apéndice III para hallar los valores de a para los que: *a*) el área a la derecha de a sea 0.05, *b*) el total del área sombreada sea 0.05, *c*) el total del área que no está sombreada sea 0.99, *d*) el área sombreada de la izquierda sea 0.01 y *e*) el área a la izquierda de a sea 0.90. Hallar los incisos del *a*) al *e*) empleando EXCEL.

SOLUCIÓN

a) Si el área sombreada a la derecha de a es 0.05, el área a la izquierda de a es $(1 - 0.05) = 0.95$, y a representa el percentil 95, $t_{.95}$. En el apéndice III, se desciende por la columna cuyo encabezado es ν hasta llegar a la entrada 9, después se avanza a la derecha hasta la columna cuyo encabezado es $t_{.95}$; el resultado, 1.83, es el valor de t que se busca.

b) Si el total del área sombreada es 0.05, entonces, por simetría, el área sombreada de la derecha es 0.025. Por lo tanto, el área a la izquierda de a es $(1 - 0.025) = 0.975$ y a representa el percentil 97.5, $t_{.975}$. En el apéndice III se encuentra que 2.26 es el valor de t buscado.

c) Si el total del área no sombreada es 0.99, entonces el total del área sombreada es $(1 - 0.99) = 0.01$ y el área sombreada a la derecha de a es $0.01/2 = 0.005$. En el apéndice III se encuentra que $t_{.995} = 3.25$.

d) Si el área sombreada a la izquierda es 0.01, entonces por simetría el área sombreada a la derecha es 0.01. En el apéndice III, $t_{.99} = 2.82$. Por lo tanto, el valor crítico de t para el cual el área sombreada a la izquierda es 0.01 es igual a -2.82.

e) Si el área sombreada a la izquierda de a es 0.90, a corresponde al percentil 90, $t_{.90}$, el cual en el apéndice III se encuentra que es igual a 1.38.

Usando EXCEL, con la expresión =TINV(0.1,9) se obtiene 1.833113. EXCEL requiere la suma de las áreas en las dos colas y los grados de libertad. De igual manera, con =TINV(0.05,9) se obtiene 2.262157, con =TINV(0.01,9) se obtiene 3.249836, con =TINV(0.02,9) se obtiene 2.821438 y con =TINV(0.2,9) se obtiene 1.383029.

11.2 Encontrar los valores críticos de t para los cuales el área de la cola derecha de la distribución t es 0.05, siendo el número de grados de libertad, ν, igual a: *a*) 16, *b*) 27 y *c*) 200.

SOLUCIÓN

Usando el apéndice III, en la columna cuyo encabezado es $t_{.95}$ se encuentran los valores: *a*) 1.75, correspondiente a $\nu = 16$; *b*) 1.70, correspondiente a $\nu = 27$ y *c*) 1.645, correspondiente a $\nu = 200$. (El último es el valor que se obtendría usando la curva normal; en el apéndice III este valor corresponde a la entrada en el último renglón marcado ∞, o infinito.)

11.3 Los coeficientes de confianza del 95% (dos colas) en la distribución normal son ± 1.96. ¿Cuáles son los coeficientes correspondientes en la distribución t para: *a*) $\nu = 9$, *b*) $\nu = 20$, *c*) $\nu = 30$ y *d*) $\nu = 60$?

SOLUCIÓN

Para los coeficientes de confianza de 95% (dos colas), el total del área sombreada en la figura 11-4 debe ser 0.05; por lo tanto, el área sombreada de la cola derecha debe ser 0.025 y el correspondiente valor de t es $t_{.975}$. Entonces, los coeficientes de confianza buscados son $\pm t_{.975}$, que para los valores de ν dados son: *a*) ± 2.26, *b*) ± 2.09, *c*) ± 2.04 y *d*) ± 2.00.

11.4 En una muestra de 10 mediciones del diámetro de una esfera, la media es $\bar{X} = 438$ centímetros (cm) y la desviación estándar es $s = 0.06$ cm. Encontrar los límites de confianza de: *a*) 95% y *b*) 99% para el verdadero diámetro.

SOLUCIÓN

a) Los límites de confianza del 95% están dados por $\bar{X} \pm t_{.975}(s/\sqrt{N-1})$.

Como $\nu = N - 1 = 10 - 1 = 9$, se encuentra que $t_{.975} = 2.26$ [ver también el problema 11.3*a*)]. Después, usando $\bar{X} = 4.38$ y $s = 0.06$, los límites de confianza buscados de 95% son $4.38 \pm 2.26(0.06/\sqrt{10-1}) = 4.38 \pm 0.0452$ cm. Por lo tanto, se puede tener una confianza de 95% en que la verdadera media se encuentra entre $(438 - 0.045) = 4.335$ cm y $(4.38 + 0.045) = 4.425$ cm.

b) Los límites de confianza del 99% están dados por $\bar{X} \pm t_{.995}(s/\sqrt{N-1})$.

Para $\nu = 9$, $t_{.995} = 3.25$. Entonces, los límites de confianza del 99% son $4.38 \pm 3.25(0.06/\sqrt{10-1}) = 4.38 \pm 0.0650$ cm y el intervalo de confianza de 99% es 4.315 a 4.445 cm.

11.5 De 25 trabajadores seleccionados en forma aleatoria se registró la cantidad de días que el año pasado faltaron al trabajo debido al síndrome del túnel carpiano, relacionado con el trabajo. Los resultados se presentan en la tabla 11.1. Cuando se usan estos datos para establecer un intervalo de confianza para la media poblacional de todos los casos, relacionados con el trabajo, de síndrome del túnel carpiano, se supone que el número de días de ausencia se distribuye normalmente en la población. Usar los datos para probar la suposición de normalidad, y si se está dispuesto a asumir la normalidad, entonces dar un intervalo de 95% para μ.

Tabla 11.1

21	23	33	32	37
40	37	29	23	29
24	32	24	46	32
17	29	26	46	27
36	38	28	33	18

SOLUCIÓN

La gráfica de probabilidad normal de MINITAB (figura 11-5) indica que la suposición de la normalidad es razonable, ya que el valor p es mayor a 0.15. Este valor p se usa para probar la hipótesis nula de que los datos han sido tomados de una población distribuida en forma normal. Empleando el nivel de significancia convencional, 0.05, la normalidad de la distribución de la población se rechazaría sólo si el valor p fuera menor a 0.05. Como se indica que el valor p correspondiente a la prueba de Kolmogorov-Smirnov para normalidad es valor $p > 0.15$, no se rechaza la suposición de normalidad.

Usando MINITAB, el intervalo de confianza que se encuentra es el siguiente. El intervalo de confianza de 95% para la media poblacional va de 27.21 a 33.59 días por año.

```
MTB > tinterval 95% confidence for data in cl
```

Intervalos de confianza

```
Variable    N     Mean     StDev     SE Mean        95.0% CI
days       25    30.40     7.72        1.54     (27.21, 33.59)
```

11.6 El espesor de las arandelas producidas con una máquina es 0.050 pulgadas (in). Para determinar si la máquina está trabajando de manera adecuada se toma una muestra de 10 arandelas en las cuales el espesor medio es 0.053 in y la desviación estándar es 0.003 in. Probar la hipótesis de que la máquina está trabajando en forma adecuada usando los niveles de significancia: *a*) 0.05 y *b*) 0.01.

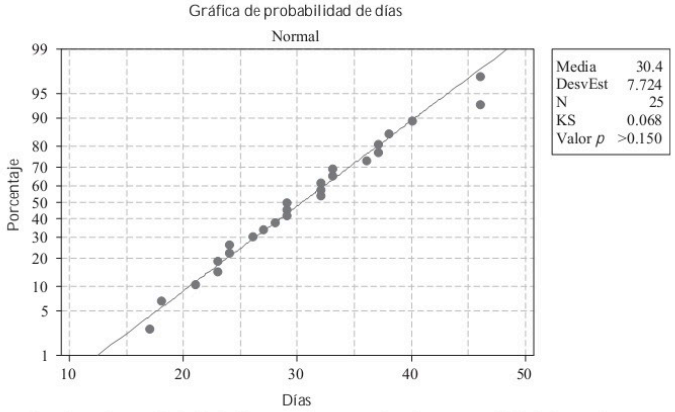


Figura 11-5 Gráfica de probabilidad normal y prueba de normalidad de Kolmogorov-Smirnov.

SOLUCIÓN

Se desea decidir entre las dos hipótesis:

H_0: $\mu = 0.050$, la máquina está trabajando de manera adecuada.
H_1: $\mu \neq 0.050$, la máquina no está trabajando en forma adecuada.

Por lo tanto, se requiere una prueba de dos colas. De acuerdo con la hipótesis H_0 se tiene

$$t = \frac{\bar{X} - \mu}{s}\sqrt{N-1} = \frac{0.053 - 0.050}{0.003}\sqrt{10-1} = 3.00$$

a) Para una prueba de dos colas a nivel de significancia 0.05, se adopta la siguiente regla de decisión:

Aceptar H_0 si t se encuentra dentro del intervalo $-t_{.975}$ a $t_{.975}$, el cual para $10 - 1 = 9$ grados de libertad es el intervalo -2.26 a 2.26.

Rechazar H_0 si no es así.

Como $t = 3.00$, se rechaza H_0 al nivel 0.05.

b) Para una prueba de dos colas al nivel de significancia 0.01, se adopta la siguiente regla de decisión:

Aceptar H_0 si t se encuentra dentro del intervalo $-t_{.995}$ a $t_{.995}$, el cual para $10 - 1 = 9$ grados de libertad es el intervalo -3.25 a 3.25.

Rechazar H_0 si no es así.

Como $t = 3.00$, se acepta H_0 al nivel de significancia 0.01.

Como H_0 se puede rechazar al nivel de significancia 0.05 pero no al nivel de significancia 0.01, se dice que la muestra da como resultado una *probabilidad significativa* (ver esta terminología al final del problema 10.5). Por lo tanto, será recomendable verificar el funcionamiento de la máquina o, por lo menos, tomar otra muestra.

11.7 El gerente de un centro comercial realiza una prueba de hipótesis para probar $\mu = \$50$ contra $\mu \neq \$50$, donde μ representa la cantidad media que gasta un comprador en ese centro comercial. En los datos que se presentan en la tabla 11.2 se dan las cantidades, en dólares, gastadas por 28 personas en el centro comercial. Para esta prueba de hipótesis, usando la distribución t de Student, se supone que los datos empleados para la prueba han sido tomados de una población distribuida normalmente. Esta suposición de normalidad puede comprobarse usando cualquiera de los métodos para *pruebas de normalidad.* MINITAB tiene tres posibilidades diferentes para pruebas de normalidad. Probar la normalidad al nivel de significancia convencional $\alpha = 0.05$. Si la suposición de normalidad no se rechaza, entonces se procede a realizar la prueba de hipótesis en que $\mu = \$50$ contra la alternativa $\mu \neq \$50$ empleando $\alpha = 0.05$.

Tabla 11.2

68	49	45	76	65	50
54	92	24	36	60	66
57	74	52	75	36	40
62	56	94	57	64	
72	65	59	45	33	

SOLUCIÓN

Empleando la prueba para normalidad de Anderson-Darling de MINITAB se obtiene el valor $p = 0.922$, la prueba de normalidad de Ryan-Joyner da un valor p mayor a 0.10, y la prueba de normalidad de Kolmogorov-Smirnov da un valor p mayor a 0.15. Al nivel de significancia convencional de 5%, en ninguno de los tres casos se puede rechazar la hipótesis de que los datos han sido tomados de una población distribuida normalmente. Recuérdese que una hipótesis nula se rechaza sólo si el valor p es menor que el nivel de significancia preestablecido. A continuación se presenta el análisis de MINITAB para la prueba de la cantidad media gastada por los clientes. Empleando el método clásico para pruebas de hipótesis, la hipótesis nula se rechaza si el valor encontrado para el estadístico de prueba es mayor, en valor absoluto, a 2.05. El valor crítico, 2.05, se encuentra empleando la distribución t de Student para 27 grados de libertad. Como el valor hallado para el estadístico de prueba es 18.50, se rechaza la hipótesis nula y se concluye que la cantidad media gastada por los clientes es mayor a \$50. Si hace la prueba de hipótesis empleando el método del valor p, entonces como el valor $p = 0.0000$ es menor al nivel de significancia (0.05), también se rechaza la hipótesis nula.

Despliegue de datos

```
Amount
68      54    57    62    72    49    92    74    56
65      45    24    52    94    59    76    36    75
57      45    65    60    36    64    33    50    66
40

MTB > TTest 0.0 'Amount';
SUBC > Alternative 0.
```

T–Test of the Mean

```
Test of mu = 0.00 vs mu not = 0.00

Variable   N     Mean     StDev     SE Mean        T         P
Amount    28    58.07     16.61        3.14    18.50    0.0000
```

11.8 El cociente intelectual (CI) de 16 estudiantes de una región de una ciudad resultó con una media de 107 y una desviación estándar de 10, el CI de 14 estudiantes de otra región de esa ciudad resultó de 112 y la desviación estándar de 8. Al nivel de significancia: *a*) 0.01 y *b*) 0.05, ¿hay diferencia entre los CI de estos dos grupos?

SOLUCIÓN

Si μ_1 y μ_2, respectivamente, denotan las medias poblacionales de los CI de los estudiantes de estas dos regiones, hay que decidir entre las hipótesis:

H_0: $\mu_1 = \mu_2$, en esencia no hay diferencia entre los dos grupos.
H_1: $\mu_1 \neq \mu_2$, hay una diferencia significativa entre los dos grupos.

De acuerdo con la hipótesis H_0,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{1/N_1 + 1/N_2}} \qquad \text{donde} \qquad \sigma = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}}$$

Por lo tanto, $$\sigma = \sqrt{\frac{16(10)^2 + 14(8)^2}{16 + 14 - 2}} = 9.44 \qquad \text{y} \qquad t = \frac{112 - 107}{9.44\sqrt{1/16 + 1/14}} = 1.45$$

a) Empleando una prueba de dos colas al nivel de significancia 0.01, H_0 se rechaza si t queda fuera del intervalo $-t_{.995}$ a $t_{.995}$, el cual para $(N_1 + N_2 - 2) = (16 + 14 - 2) = 28$ grados de libertad es el intervalo -2.76 a 2.76. Por lo tanto, al nivel de significancia 0.01 no se puede rechazar H_0.

b) Empleando una prueba de dos colas al nivel de significancia 0.05, H_0 se rechaza si t queda fuera del intervalo $-t_{.975}$ a $t_{.975}$, el cual para 28 grados de libertad es el intervalo -2.05 a 2.05. Por lo tanto, al nivel de significancia 0.05 no se puede rechazar H_0.

Se concluye que no hay una diferencia significativa entre los CI de los dos grupos.

11.9 En la tabla 11.3 se dan los costos anuales (en miles de dólares) de colegiatura, alojamiento y manutención en 10 universidades privadas elegidas en forma aleatoria y 15 universidades públicas elegidas en forma aleatoria. Probar la hipótesis nula de que el costo medio anual en las universidades privadas es 10 mil dólares mayor al costo medio anual en las universidades públicas, contra la hipótesis alternativa de que la diferencia no es de 10 mil dólares. Usar el nivel de significancia 0.05. Antes de realizar la prueba de las medias, probar, al nivel de significancia 0.05, la suposición de normalidad y de varianzas iguales.

Tabla 11.3

Universidades públicas			Universidades privadas	
4.2	9.1	11.6	13.0	17.7
6.1	7.7	10.4	18.8	17.6
4.9	6.5	5.0	13.2	19.8
8.5	6.2	10.4	14.4	16.8
4.6	10.2	8.1	17.7	16.1

SOLUCIÓN

En la figura 11-6 se muestran los resultados de MINITAB para la prueba de normalidad de Anderson-Darling de las universidades públicas. Dado que el valor p (0.432) no es menor a 0.05, la suposición de normalidad no se rechaza. Una

prueba similar para las universidades privadas indica que la suposición de normalidad también es válida para las universidades privadas.

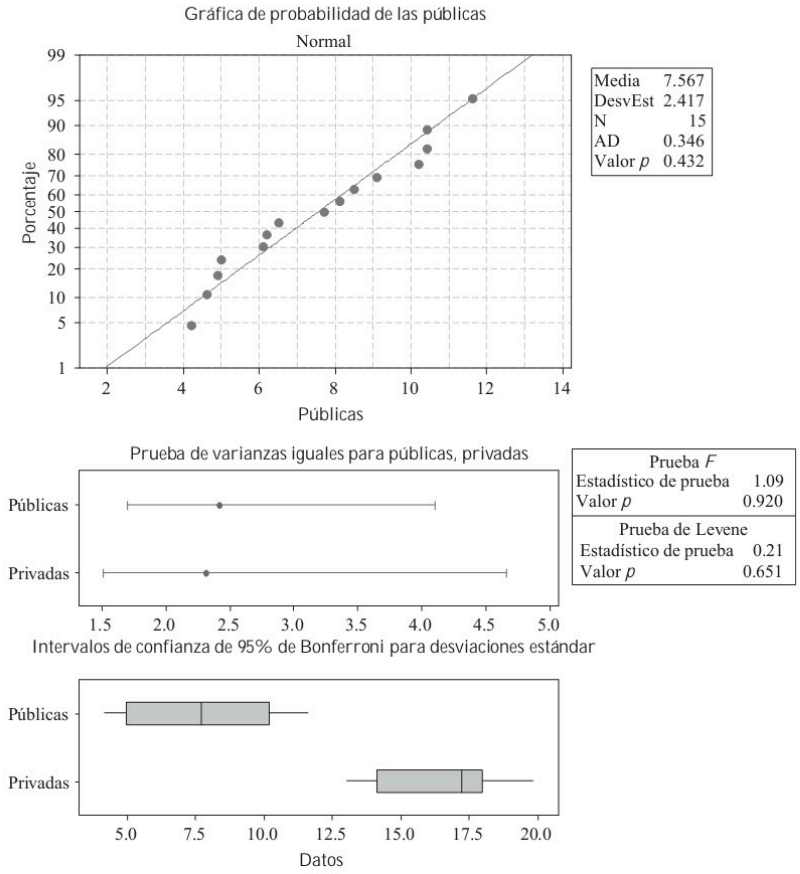


Figura 11-6 Prueba de normalidad de Anderson-Darling y prueba *F* de varianzas iguales.

La prueba *F* que se muestra en la parte inferior de la figura 11-6 indica que puede suponerse que las varianzas son iguales. Con la secuencia de comandos **"Stat ⇒ Basic Statistics ⇒ 2-sample t"** se obtiene el resultado que se da a continuación. Los resultados indican que no se puede rechazar que el costo de las universidades privadas sea 10 mil dólares mayor al de las universidades públicas.

Prueba T de dos muestras y CI: públicas, privadas

```
Two-sample T for Public vs Private

            N        Mean       StDev     SE Mean
Public     15        7.57        2.42        0.62
Private    10       16.51        2.31        0.73

Difference = mu (Public) - mu(Private)
Estimate for difference: -8.9433
95% CI for difference: (-10.9499, -6.9367)
T-Test of difference = -10 (vs not =) : T-Value = 1.09 P-Value = 0.287
DF = 23
Both use Pooled StDev = 2.3760
```

DISTRIBUCIÓN JI CUADRADA

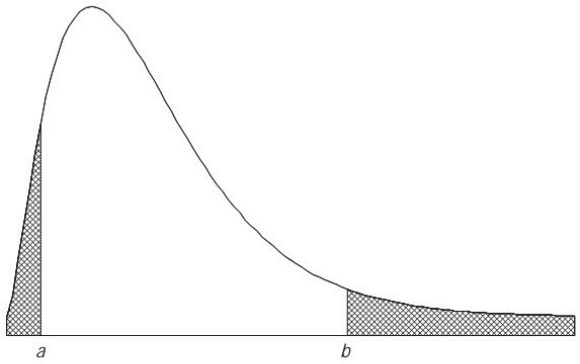


Figura 11-7 Distribución ji cuadrada para 5 grados de libertad.

11.10 En la figura 11-7 se muestra la gráfica de la distribución ji cuadrada para 5 grados de libertad. Empleando el apéndice IV, hallar los valores críticos de χ^2 para los cuales: *a*) el área sombreada de la derecha es 0.05, *b*) el total del área sombreada es 0.05, *c*) el área sombreada de la izquierda es 0.10 y *d*) el área sombreada de la derecha es 0.01. Hallar también estas respuestas usando EXCEL.

SOLUCIÓN

a) Si el área sombreada de la derecha es 0.05, entonces el área a la izquierda de *b* es $(1 - 0.05) = 0.95$ y *b* es el percentil 95, $\chi^2_{.95}$. Refiérase al apéndice IV, bajar por la columna que tiene como encabezado ν hasta llegar a la entrada 5, y después avanzar hacia la derecha hasta la columna cuyo encabezado es $\chi^2_{.95}$; el resultado, 11.1, es el valor crítico de χ^2 que se busca.

b) Como esta distribución no es simétrica, hay muchos valores críticos para los que el total del área sombreada es 0.05. Por ejemplo, el área sombreada de la derecha puede ser 0.04 y el área sombreada de la izquierda 0.01. Sin embargo, se acostumbra, a menos que se especifique otra cosa, elegir estas áreas de manera que sean iguales. En este caso, entonces, cada área es 0.025. Si el área sombreada de la derecha es 0.025, el área a la izquierda de *b* es $1 - 0.025 = 0.975$ y *b* es el percentil 97.5, $\chi^2_{.975}$, el cual de acuerdo con el apéndice IV es 12.8. De igual manera, si el área sombreada de la izquierda es 0.025, el área a la izquierda de *a* es 0.025 y *a* es el percentil 2.5, $\chi^2_{.025}$, que es igual a 0.831. Por lo tanto, los valores críticos son 0.83 y 12.8.

c) Si el área sombreada de la izquierda es 0.10, *a* representa el percentil 10, $\chi^2_{.10}$, el cual es igual a 1.61.

d) Si el área sombreada de la derecha es 0.01, el área a la izquierda de *b* es 0.99 y *b* representa el percentil 99, $\chi^2_{.99}$, el cual es igual a 15.1.

La respuesta de EXCEL para *a*) se obtiene con =CHIINV(0.05,5), que da 11.0705. El primer parámetro de CHIINV es el área a la derecha del punto y el segundo es el número de grados de libertad. La respuesta para *b*) se obtiene con =CHIINV(0.975,5), que da 0.8312 y =CHIINV(0.025,5) da 12.8325. La respuesta para *c*) se obtiene con =CHIINV(0.9,5), que da 1.6103. La respuesta para *d*) se obtiene con =CHIINV(0.01,5), que da 15.0863.

11.11 Encontrar el valor crítico de χ^2 tal que el área en la cola derecha de la distribución χ^2 sea 0.05, siendo el número de grados de libertad, ν, igual a: *a*) 15, *b*) 21 y *c*) 50.

SOLUCIÓN

En el apéndice IV, en la columna cuyo encabezado es $\chi^2_{.95}$ se encuentran los valores: *a*) 25.0 que corresponde a $\nu = 15$; *b*) 32.7 que corresponde a $\nu = 21$ y *c*) 67.5 que corresponde a $\nu = 50$.

11.12 Encontrar el valor mediano de χ^2 que corresponda a: *a*) 9, *b*) 28 y *c*) 40 grados de libertad.

SOLUCIÓN

En el apéndice IV, en la columna cuyo encabezado es $\chi^2_{.50}$ (ya que la mediana es el percentil 50), se encuentran los valores: *a*) 8.34, que corresponde a $\nu = 9$; *b*) 27.3, que corresponde a $\nu = 28$, y *c*) 39.3, que corresponde a $\nu = 40$.

Resulta interesante observar que los valores medianos están muy cercanos a la igualdad del número de grados de libertad. De hecho, para $\nu > 10$, los valores medianos son iguales a $(\nu - 0.7)$, como puede verse en la tabla.

11.13 La desviación estándar de las estaturas de 16 estudiantes elegidos en forma aleatoria en una escuela de 1 000 estudiantes es 2.40 in. Encontrar los límites de confianza de: *a*) 95% y *b*) 99% para la desviación estándar de las estaturas de todos los estudiantes de esta escuela.

SOLUCIÓN

a) Los límites de confianza de 95% son $s\sqrt{N}/\chi_{.975}$ y $s\sqrt{N}/\chi_{.025}$.

Para $\nu = 16 - 1 = 15$ grados de libertad, $\chi^2_{.975} = 27.5$ (o bien $\chi_{.975} = 5.24$) y $\chi^2_{.025} = 6.26$ (o bien $\chi_{.025} = 2.50$). Los límites de confianza de 95% son $2.40\sqrt{16}/5.24$ y $2.40\sqrt{16}/2.50$ (es decir, 1.83 y 3.84 in). Por lo tanto, se puede tener una confianza de 95% de que la desviación estándar poblacional se encuentra entre 1.83 y 3.84 in.

b) Los límites de confianza de 99% son $s\sqrt{N}/\chi_{.995}$ y $s/\sqrt{N}/\chi_{.005}$.

Para $\nu = 16 - 1 = 15$ grados de libertad están dados por $\chi^2_{.995} = 32.8$ (o $\chi_{.995} = 5.73$) y $\chi^2_{.005} = 4.60$ (o bien $\chi_{.005} = 2.14$). Entonces los límites de confianza de 99% son $2.40\sqrt{16}/5.73$ y $2.40\sqrt{16}/2.14$ (es decir, 1.68 y 4.49 in). Por lo tanto, se puede tener una confianza de 99% de que la desviación estándar poblacional se encuentre entre 1.68 y 4.49 in.

11.14 Encontrar $\chi^2_{.95}$ para: *a*) $\nu = 50$ y *b*) $\nu = 100$ grados de libertad.

SOLUCIÓN

Para ν mayor que 30 se puede emplear el hecho de que $\sqrt{2\chi^2} - \sqrt{2\nu - 1}$ es una distribución aproximadamente normal en la que la media es 0 y la desviación estándar es 1. Entonces, si z_p es un percentil de la puntuación z en la distribución normal estándar, se puede escribir, con un alto grado de aproximación,

$$\sqrt{2\chi^2_p} - \sqrt{2\nu - 1} = z_p \qquad \text{o} \qquad \sqrt{2\chi^2_p} = z_p + \sqrt{2\nu - 1}$$

de donde $\chi^2_p = \frac{1}{2}(z_p + \sqrt{2\nu - 1})^2$.

a) Si $\nu = 50$, $\chi^2_{.95} = \frac{1}{2}(z_{.95} + \sqrt{2(50) - 1})^2 = \frac{1}{2}(1.64 + \sqrt{99})^2 = 67.2$, lo que coincide muy bien con el valor 67.5 dado en el apéndice IV.

b) Si $\nu = 100$, $\chi^2_{.95} = \frac{1}{2}(z_{.95} + \sqrt{2(100) - 1})^2 = \frac{1}{2}(1.64 + \sqrt{199})^2 = 124.0$ (verdadero valor = 124.3).

11.15 La desviación estándar del tiempo de vida de una muestra de 200 bombillas eléctricas es 100 horas (h). Encontrar los límites de confianza de: *a*) 95% y *b*) 99% para la desviación estándar de estas bombillas eléctricas.

SOLUCIÓN

a) Los límites de confianza de 95% están dados por $s\sqrt{N}/\chi_{.975}$ y $s\sqrt{N}/\chi_{.025}$.

Para $\nu = 200 - 1 = 199$ grados de libertad, se encuentra (como en el problema 11.14)

$$\chi^2_{.975} = \tfrac{1}{2}(z_{.975} + \sqrt{2(199) - 1})^2 = \tfrac{1}{2}(1.96 + 19.92)^2 = 239$$
$$\chi^2_{.025} = \tfrac{1}{2}(z_{.025} + \sqrt{2(199) - 1})^2 = \tfrac{1}{2}(-1.96 + 19.92)^2 = 161$$

por lo tanto, $\chi_{.975} = 15.5$ y $\chi_{0.025} = 12.7$. De manera que los límites de confianza del 95% son $100\sqrt{200}/15.5 = 91.2$ h y $100\sqrt{200}/12.7 = 111.3$ h, respectivamente. Se puede tener una confianza de 95% en que la desviación estándar poblacional esté entre 91.2 y 111.3 h.

b) Los límites de confianza de 99% están dados por $s\sqrt{N}/\chi_{.995}$ y $s\sqrt{N}/\chi_{.005}$.
Para $\nu = 200 - 1 = 199$ grados de libertad,

$$\chi^2_{.995} = \tfrac{1}{2}(z_{.995} + \sqrt{2(199) - 1})^2 = \tfrac{1}{2}(2.58 + 19.92)^2 = 253$$
$$\chi^2_{.005} = \tfrac{1}{2}(z_{.005} + \sqrt{2(199) - 1})^2 = \tfrac{1}{2}(-2.58 + 19.92)^2 = 150$$

por lo tanto, $\chi_{.995} = 15.9$ y $\chi_{.005} = 12.2$. De manera que los límites de confianza del 99% son $100\sqrt{200}/15.9 = 88.9$ h y $100\sqrt{200}/12.2 = 115.9$ h, respectivamente. Se puede tener una confianza de 99% en que la desviación estándar poblacional esté entre 88.9 y 115.9 h.

11.16 Un fabricante de ejes requiere que en el proceso de fabricación el diámetro de los ejes sea 5.000 cm. Además, para garantizar que las ruedas se ajusten de manera adecuada a los ejes, es necesario que la desviación estándar en los diámetros sea 0.005 cm o menos. En la tabla 11.4 se presentan los diámetros de los 20 ejes de una muestra.

Tabla 11.4

4.996	4.998	5.002	4.999
5.010	4.997	5.003	4.998
5.006	5.004	5.000	4.993
5.002	4.996	5.005	4.992
5.007	5.003	5.000	5.000

El fabricante desea probar la hipótesis nula de que la desviación estándar poblacional es 0.005 cm contra la hipótesis alternativa de que la desviación estándar poblacional es mayor a 0.005 cm. Si se confirma la hipótesis alternativa, entonces el proceso de fabricación debe detenerse y deben hacerse ajustes a las máquinas. Para la prueba se supone que los diámetros de los ejes tienen una distribución normal. Probar esta suposición al nivel de significancia 0.05. Si se está dispuesto a suponer normalidad, entonces hacer la prueba concerniente a la desviación estándar poblacional al nivel de significancia 0.05.

SOLUCIÓN

En la figura 11-8 se muestra la prueba de normalidad de Shapiro-Wilk. Como el valor *p* que se obtiene es grande (0.9966), no se puede rechazar la normalidad. Esta gráfica de probabilidad y el análisis de Shapiro-Wilk se hicieron empleando el paquete STATISTIX de software para estadística.

Se tiene que decidir entre las hipótesis:

H_0: $\sigma = 0.005$ cm, el valor observado se debe a la casualidad.
H_1: $\sigma = 0.005$ cm, la variabilidad es demasiado grande.

El análisis realizado con SAS es el siguiente:

```
One Sample Chi-square Test for a Variance
Sample Statistics for diameter
N                  Mean       Std. Dev.            Variance
-------------------------------------------------------------
20                5.0006        0.0046              215E-7

Hypothesis Test
Null hypothesis: Variance of diameter <=0.000025
Alternative: Variance of diameter > 0.000025
Chi-square          Df           Prob
-------------------------------------
16.358              19          0.6333
```

Como el valor *p* obtenido (0.6333) es grande, esto indica que la hipótesis nula no se debe rechazar.

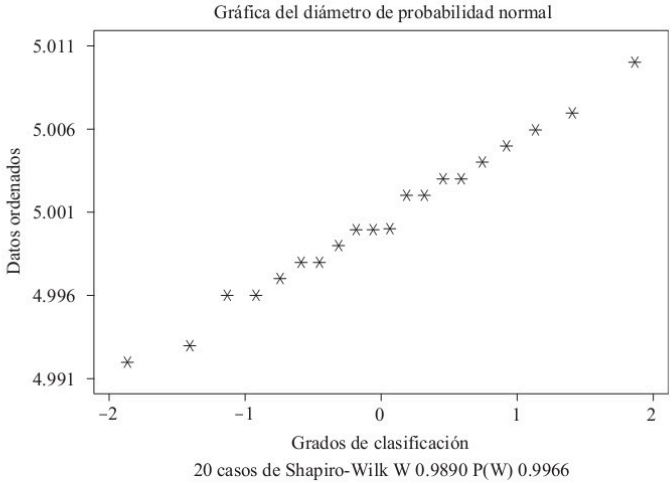


Figura 11-8 STATISTIX, prueba de normalidad de Shapiro-Wilk.

11.17 La desviación estándar en los pesos de paquetes de 40.0 onzas (oz), llenados con una máquina, ha sido 0.25 oz. En una muestra de 20 paquetes se observa una desviación estándar de 0.32 oz. ¿Este aparente incremento en la variabilidad es significativo a los niveles: *a*) 0.05 y *b*) 0.01?

SOLUCIÓN

Decidir entre las hipótesis:

$H_0 : \sigma = 0.25$ oz, el resultado observado es casualidad.
$H_1 : \sigma > 0.25$ oz, la variabilidad ha aumentado.

El valor de χ^2 para la muestra es

$$\chi^2 = \frac{Ns^2}{\sigma^2} = \frac{20(0.32)^2}{(0.25)^2} = 32.8$$

a) Empleando una prueba de una cola, al nivel de significancia 0.05, se rechaza H_0 si los valores muestrales de χ^2 son mayores a $\chi^2_{.95}$, lo que es igual a 30.1 para $\nu = 20 - 1 = 19$ grados de libertad. Por lo tanto, se rechaza H_0 al nivel de significancia 0.05.

b) Empleando una prueba de una cola, al nivel de significancia 0.01, se puede rechazar H_0 si los valores muestrales de χ^2 son mayores a $\chi^2_{.99}$, lo que es igual a 36.2 para 19 grados de libertad. Por lo tanto, al nivel de significancia 0.01, no se rechaza H_0.

Se concluye que la variabilidad probablemente ha aumentado. Se recomienda examinar la máquina.

LA DISTRIBUCIÓN *F*

11.18 De poblaciones distribuidas en forma normal se obtienen dos muestras de tamaños 9 y 12 cuyas varianzas son 16 y 25. Si las varianzas muestrales son 20 y 8, respectivamente, determinar si la primera muestra tiene una varianza bastante mayor que la segunda muestra al nivel de significancia: *a*) 0.05, *b*) 0.01 y *c*) usar EXCEL para mostrar que el área a la derecha de 4.03 está entre 0.01 y 0.05.

SOLUCIÓN

Para estas dos muestras, 1 y 2, se tiene $N_1 = 9$, $N_2 = 12$, $\sigma_1^2 = 16$, $\sigma_2^2 = 25$, $S_1^2 = 20$ y $S_2^2 = 8$. Por lo tanto,

$$F = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2} = \frac{N_1 S_1^2/(N_1 - 1)\sigma_1^2}{N_2 S_2^2/(N_2 - 1)\sigma_2^2} = \frac{(9)(20)/(9-1)(16)}{(12)(8)/(12-1)(25)} = 4.03$$

a) Los grados de libertad para el numerador y para el denominador de F son $\nu_1 = N_1 - 1 = 9 - 1 = 8$ y $\nu_2 = N_2 - 1 = 12 - 1 = 11$. Entonces, en el apéndice V se encuentra que $F_{.95} = 2.95$. Como el valor de F calculado es $F = 4.03$, que es mayor a 2.95, se concluye que la varianza de la muestra 1 es significativamente mayor que la de la muestra 2, al nivel de significancia 0.05.

b) Para $\nu_1 = 8$ y $\nu_2 = 11$, en el apéndice VI se encuentra $F_{.01} = 4.74$. En este caso el valor de F calculado es $F = 4.03$, que es menor a 4.74. Por lo tanto, no se puede concluir que la varianza de la muestra 1 sea mayor que la varianza de la muestra 2, al nivel de significancia 0.01.

c) El área a la derecha de 4.03 está dada por =FDIST(4.03,8,11) y es 0.018.

11.19 De dos poblaciones distribuidas de manera normal se toman dos muestras, una de tamaño 8 y otra de tamaño 10, cuyas varianzas corresponden a 20 y 36. Encontrar la probabilidad de que la varianza de la primera muestra sea mayor al doble de la varianza de la segunda muestra.

Usar EXCEL para hallar la probabilidad exacta de que F con 7 y 9 grados de libertad sea mayor a 3.70.

SOLUCIÓN

Se tiene $N_1 = 8$, $N_2 = 10$, $\sigma_1^2 = 20$, y $\sigma_2^2 = 36$. Por lo tanto,

$$F = \frac{8S_1^2/(7)(20)}{10S_2^2/(9)(36)} = 1.85\frac{S_1^2}{S_2^2}$$

El número de grados de libertad en el numerador y en el denominador es $\nu_1 = N_1 - 1 = 8 - 1 = 7$ y $\nu_2 = N_2 - 1 = 10 - 1 = 9$. Ahora, si S_1^2 es mayor al doble de S_2^2, entonces

$$F = 1.85\frac{S_1^2}{S_2^2} > (1.85)(2) = 3.70$$

Buscando 3.70 en los apéndices V y VI se encuentra que la probabilidad es menor a 0.05 pero mayor a 0.01. Para encontrar los valores exactos se necesita una tabulación más extensa que la distribución F.

Con EXCEL la respuesta se obtiene con =FDIST(3.7,7,9), que da 0.036, que es la probabilidad de que F con 7 y 9 grados de libertad sea mayor a 3.70.

PROBLEMAS SUPLEMENTARIOS

DISTRIBUCIÓN *t* DE STUDENT

11.20 En una distribución de Student con 15 grados de libertad, encontrar el valor de t_1 tal que: *a*) el área a la derecha de t_1 sea 0.01, *b*) el área a la izquierda de t_1 sea 0.95, *c*) el área a la derecha de t_1 sea 0.10, *d*) el área a la derecha de t_1 junto con el área a la izquierda de $-t_1$ sea 0.01 y *e*) el área entre $-t_1$ y t_1 sea 0.95.

11.21 Usando el apéndice III, encontrar los valores críticos de t para los cuales el área en la cola derecha de la distribución t sea 0.01, siendo el número de grados de libertad, ν, igual a: *a*) 4, *b*) 12, *c*) 25, *d*) 60 y *e*) 150. Dar las soluciones de *a*) a *e*) usando EXCEL.

11.22 En la distribución t de Student encontrar los valores de t_1 que satisfacen cada una de las condiciones siguientes:

a) El área entre $-t_1$ y t_1 es 0.90 y $\nu = 25$.
b) El área a la izquierda de $-t_1$ es 0.025 y $\nu = 20$.
c) El área a la derecha de t_1 junto con el área a la izquierda de $-t_1$ es 0.01 y $\nu = 5$.
d) El área a la derecha de t_1 es 0.55 y $\nu = 16$.

11.23 Si una variable U tiene una distribución t de Student con $\nu = 10$, encontrar la constante C que satisfaga: *a*) $\Pr\{U > C\} = 0.05$, *b*) $\Pr\{-C \leq U \leq C\} = 0.98$, *c*) $\Pr\{U \leq C\} = 0.20$ y *d*) $\Pr\{U \geq C\} = 0.90$.

11.24 En la distribución normal, los coeficientes de confianza de 99% (dos colas) son ± 2.58. ¿Cuáles son los coeficientes correspondientes en la distribución t si: *a*) $\nu = 4$, *b*) $\nu = 12$, *c*) $\nu = 25$, *d*) $\nu = 30$, y *e*) $\nu = 40$?

11.25 En una muestra de 12 mediciones de la resistencia a la ruptura de un hilo de algodón, la media es 7.38 gramos (g) y la desviación estándar 1.24 g. Encontrar los límites de confianza de: *a*) 95% y *b*) 99% para la verdadera resistencia a la ruptura y *c*) la solución que da MINITAB usando el resumen de estadísticos.

11.26 Resolver el ejercicio 11.25 suponiendo que los métodos de la teoría de muestras grandes son aplicables, y comparar los resultados obtenidos.

11.27 Se tomaron cinco mediciones del tiempo de reacción de una persona a cierto estímulo; las mediciones fueron 0.28, 0.30, 0.27, 0.33 y 0.31 segundos. Encontrar los límites de confianza de: *a*) 95% y *b*) 99% para el verdadero tiempo de reacción.

11.28 El tiempo medio de vida de los focos eléctricos producidos por una empresa ha sido 1 120 h y la desviación estándar 125 h. En una muestra de 8 focos eléctricos, recientemente producidos, el tiempo medio de vida fue de 1 070 h. Probar la hipótesis de que el tiempo medio de vida de los focos no ha variado, usando los niveles de significancia: *a*) 0.05 y *b*) 0.01.

11.29 En el problema 11.28 probar las hipótesis $\mu = 1\,120$ h contra $\mu < 1\,120$ h, usando como niveles de significancia: *a*) 0.05 y *b*) 0.01.

11.30 Las especificaciones en la producción de cierta aleación exigen 23.2% de cobre. En una muestra consistente en 10 análisis del producto, el contenido medio de cobre fue 23.5% y la desviación estándar 0.24%. A los niveles de significancia: *a*) 0.05 y *b*) 0.01 ¿puede concluirse que el producto satisface las especificaciones?

11.31 En el problema 11.30, empleando los niveles de significancia: *a*) 0.01 y *b*) 0.05, probar la hipótesis de que el contenido medio de cobre es mayor que el requerido por las especificaciones.

11.32 Un experto asegura que introduciendo un nuevo tipo de máquina en un proceso de producción se puede disminuir notablemente el tiempo de producción. Debido a los gastos requeridos para el mantenimiento de esta máquina, el gerente encuentra que a menos que el tiempo de producción se reduzca por lo menos en 8%, no vale la pena introducir la nueva máquina. Seis experimentos resultantes mostraron que el tiempo de producción se redujo en 8.4% con una desviación estándar de 0.32%. Usando como niveles de significancia: *a*) 0.01 y *b*) 0.05, probar la hipótesis de que debe introducirse la nueva máquina.

11.33 Empleando una marca A de gasolina el rendimiento medio en millas por galón encontrado en cinco automóviles similares bajo condiciones idénticas es 22.6 y la desviación estándar es 0.48. Empleando la marca B, el rendimiento medio es 21.4 y la desviación estándar es 0.54. Usando el nivel de significancia 0.05, investigar si la marca A da realmente un mejor rendimiento que la marca B.

11.34 Se prueba el pH (grado de acidez de una solución) de dos soluciones químicas, *A* y *B*. En seis muestras de *A* la media en el pH es 7.2 y la desviación estándar es 0.024. En cinco muestras de la solución *B* la media en el pH es 7.49 y la desviación estándar es 0.032. Al nivel de significancia 0.05, determinar si el pH de estos dos tipos de soluciones es diferente.

11.35 En un examen de psicología, la media de las calificaciones de los 12 estudiantes de un grupo es 78 y la desviación estándar es 6; la media de las calificaciones de los 15 estudiantes de otro grupo es 74 y la desviación estándar es 8. Empleando el nivel de significancia 0.05, determinar si el primer grupo es mejor que el segundo grupo.

LA DISTRIBUCIÓN JI CUADRADA

11.36 En el apéndice IV, en la distribución ji cuadrada para 12 grados de libertad, hallar el valor de χ_c^2 tal que: *a*) el área a la derecha de χ_c^2 sea 0.05, *b*) el área a la izquierda de χ_c^2 sea 0.99, *c*) el área a la derecha de χ_c^2 sea 0.025 y *d*) resolver los incisos del *a*) al *c*) empleando EXCEL.

11.37 Hallar los valores críticos de χ^2 para los cuales el área en la cola derecha de la distribución es 0.05, siendo el número de grados de libertad, ν, igual a: *a*) 8, *b*) 19, *c*) 29 y *d*) 40.

11.38 Resolver el problema 11.37 si el área en la cola derecha es 0.01.

11.39 *a*) Encontrar χ_1^2 y χ_2^2 tales que el área bajo la distribución χ^2 para $\nu = 20$ entre χ_1^2 y χ_2^2 sea 0.95, suponiendo áreas iguales a la derecha de χ_2^2 y a la izquierda de χ_1^2.

b) Mostrar que si en *a*) no se hace la suposición de áreas iguales, los valores χ_1^2 y χ_2^2 no son únicos.

11.40 Si una variable U tiene la distribución ji cuadrada con $\nu = 7$, encontrar χ_1^2 y χ_2^2 tales que: *a*) $\Pr\{U > \chi_2^2\} = 0.025$, *b*) $\Pr\{U < \chi_1^2\} = 0.50$ y *c*) $\Pr\{\chi_1^2 \leq U \leq \chi_2^2\} = 0.90$.

11.41 La desviación estándar encontrada en la duración de 10 bombillas eléctricas producidas por una empresa es 120 h. Encontrar los límites de confianza de: *a*) 95% y *b*) 99% para la desviación estándar de todas las bombillas eléctricas fabricadas por la empresa.

11.42 Resolver el problema 11.41 si se tienen 25 bombillas eléctricas en las que la desviación estándar es 120 h.

11.43 Encontrar: *a*) $\chi_{.05}^2$ y *b*) $\chi_{.95}^2$ para $\nu = 150$ empleando $\chi_p^2 = \frac{1}{2}(z_p + \sqrt{2v-1})^2$ y *c*) comparar estos resultados con los que se obtienen usando EXCEL.

11.44 Encontrar: *a*) $\chi_{.025}^2$ y *b*) $\chi_{.975}^2$ para $\nu = 250$ empleando $\chi_p^2 = \frac{1}{2}(z_p + \sqrt{2v-1})^2$ y *c*) comparar estos resultados con los que se obtienen usando EXCEL.

11.45 Mostrar que si se tienen valores grandes de ν, una buena aproximación a χ^2 es la dada por $(v + z_p\sqrt{2\nu})$, donde z_p es el percentil p de la distribución normal estándar.

11.46 Resolver el problema 11.39 usando la distribución χ^2 si en una muestra de 100 bombillas eléctricas se encuentra la misma desviación estándar de 120 h. Comparar los resultados con los obtenidos con los métodos del capítulo 9.

11.47 En el problema 11.44, ¿cuál es el intervalo de confianza de 95% que tiene la menor amplitud?

11.48 La desviación estándar en la resistencia a la ruptura de determinados cables producidos por una empresa es de 240 libras (lb). Después de que se introdujo una modificación en el proceso de fabricación de estos cables, en una muestra de ocho cables la desviación estándar encontrada fue 300 lb. Investigar la significancia del aparente aumento de variabilidad a los niveles de significancia: *a*) 0.05 y *b*) 0.01.

11.49 La desviación estándar de la temperatura anual de una ciudad durante 100 años fue de 16° Fahrenheit. Usando la temperatura media del día 15 de cada mes durante los últimos 15 años, la desviación estándar calculada de la temperatura anual fue de 10° Fahrenheit. Probar la hipótesis de que la temperatura en la ciudad se volvió menos variable que en el pasado, usando los niveles de significancia de: *a*) 0.05 y *b*) 0.01.

LA DISTRIBUCIÓN *F*

11.50 Empleando los apéndices V y VI, encontrar los valores de *F* que se piden en los incisos del *a*) al *d*).

a) $F_{0.95}$ para $V_1 = 8$ y $V_2 = 10$.
b) $F_{0.99}$ para $V_1 = 24$ y $V_2 = 11$.
c) $F_{0.85}$ para $N_1 = 16$ y $N_2 = 25$.
d) $F_{0.90}$ para $N_1 = 21$ y $N_2 = 23$.

11.51 Resolver el problema 11.50 usando EXCEL.

11.52 De poblaciones distribuidas normalmente cuyas varianzas son 40 y 60 se toman dos muestras de tamaños 10 y 15, respectivamente. Si las varianzas muestrales son 90 y 50, determinar si la varianza de la muestra 1 es significativamente mayor que la de la muestra 2 a los niveles de: *a*) 0.05 y *b*) 0.01.

11.53 Dos empresas, *A* y *B*, fabrican bombillas eléctricas. Los tiempos de vida de estas bombillas están distribuidos casi en forma normal y sus desviaciones estándar son 20 y 27 h, respectivamente. Si se toman 16 bombillas de la empresa *A* y 20 bombillas de la empresa *B* y se determina que las desviaciones estándar de sus tiempos de vida corresponden a 15 y 40 h, ¿puede determinarse, a los niveles de significancia: *a*) 0.05 y *b*) 0.01, que la variabilidad en las bombillas de *A* es mayor que la variabilidad en las bombillas de *B*?

LA PRUEBA JI CUADRADA

12

FRECUENCIAS OBSERVADAS Y FRECUENCIAS TEÓRICAS

Como se ha visto, los resultados obtenidos de las muestras no siempre coinciden exactamente con los resultados teóricos esperados según las reglas de la probabilidad. Por ejemplo, aunque de acuerdo con las consideraciones teóricas en 100 lanzamientos de una moneda se esperarían 50 caras y 50 cruces, es raro que se obtengan exactamente estos resultados.

Supóngase que en una muestra determinada se observa la ocurrencia de un conjunto de eventos $E_1, E_2, E_3, \ldots, E_k$ (ver tabla 12.1) con las frecuencias $o_1, o_2, o_3, \ldots, o_k$, llamadas *frecuencias observadas* y que, según las reglas de la probabilidad, se esperaría que estos eventos ocurrieran con frecuencias $e_1, e_2, e_3, \ldots, e_k$, llamadas *frecuencias esperadas* o *teóricas*. Se desea saber si las frecuencias observadas difieren, de manera significativa, de las frecuencias esperadas.

Tabla 12.1

Eventos	E_1	E_2	E_3	$\cdots$	E_k
Frecuencias observadas	o_1	o_2	o_3	$\cdots$	o_k
Frecuencias esperadas	e_1	e_2	e_3	$\cdots$	e_k

DEFINICIÓN DE χ^2

Una medida de la discrepancia entre las frecuencias observadas y las frecuencias esperadas la proporciona el estadístico χ^2 (léase ji cuadrada) dado por

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \cdots + \frac{(o_k - e_k)^2}{e_k} = \sum_{j=1}^{k} \frac{(o_j - e_j)^2}{e_j} \tag{1}$$

Donde, si la frecuencia total es N,

$$\sum o_j = \sum e_j = N \tag{2}$$

Una expresión equivalente a la fórmula (*1*) es (ver problema 12.11)

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N \tag{3}$$

Si $\chi^2 = 0$, las frecuencias observadas y las frecuencias teóricas coinciden exactamente; en tanto que si $\chi^2 > 0$, la coincidencia no es exacta. Cuanto mayor sea el valor de χ^2, mayor la discrepancia entre frecuencias observadas y frecuencias esperadas.

La distribución muestral de χ^2 se puede aproximar con bastante exactitud mediante la distribución ji cuadrada

$$Y = Y_0(\chi^2)^{1/2(\nu-2)}e^{-1/2\chi^2} = Y_0\chi^{\nu-2}e^{-1/2\chi^2} \tag{4}$$

(vista en el capítulo 11) si las frecuencias esperadas son mayores o iguales a 5. La aproximación mejora cuanto mayores sean estos valores.

El número de grados de libertad, ν, es

1. $\nu = k - 1$ si las frecuencias esperadas pueden calcularse sin tener que estimar parámetros poblacionales a partir de estadísticos muestrales. Obsérvese que a k se le resta 1 debido a la condición restrictiva (*2*), que establece que conociendo $k - 1$ de las frecuencias esperadas, queda determinada la frecuencia restante.
2. $\nu = k - 1 - m$ si las frecuencias esperadas sólo pueden calcularse estimando m parámetros poblacionales a partir de estadísticos muestrales.

PRUEBAS DE SIGNIFICANCIA

En la práctica, las frecuencias esperadas se calculan basándose en la hipótesis H_0. Si de acuerdo con esta hipótesis el valor calculado para χ^2, mediante las ecuaciones (*1*) o (*3*) es mayor a algún valor crítico (por ejemplo, $\chi^2_{.95}$ o $\chi^2_{.99}$, que son los valores críticos para los niveles de significancia 0.05 y 0.01, respectivamente), se concluye que las frecuencias observadas difieren *en forma significativa* de las frecuencias esperadas y se rechaza H_0 al correspondiente nivel de significancia; si no es así, se acepta H_0 (o por lo menos no se rechaza). A este procedimiento se le conoce como *prueba ji cuadrada* de hipótesis o de significancia.

Es necesario notar que hay que tener desconfianza de aquellas circunstancias en las que χ^2 tenga un valor *demasiado cercano a cero*, pues es raro que exista una coincidencia *tan buena* entre las frecuencias observadas y las frecuencias esperadas. Para examinar tales situaciones se determina si el valor obtenido para χ^2 es menor a $\chi^2_{.05}$ o a $\chi^2_{.01}$, en cuyo caso se decide que a los niveles de significancia 0.05 o 0.01, respectivamente, la coincidencia es *demasiado buena*.

LA PRUEBA JI CUADRADA DE BONDAD DE AJUSTE

La prueba chi cuadrada puede emplearse para determinar qué tan bien se ajustan una distribución teórica (por ejemplo, la distribución normal o la distribución binomial) a una distribución empírica (es decir, a una distribución obtenida a partir de datos muestrales). Ver los problemas 12.12 y 12.13.

EJEMPLO 1 Un par de dados se lanzan 500 veces y las sumas de las caras que caen hacia arriba son las que se muestran en la tabla 12.2.

Tabla 12.2

Suma	2	3	4	5	6	7	8	9	10	11	12
Observada	15	35	49	58	65	76	72	60	35	29	6

Los números esperados, si el dado no está cargado, se determinan a partir de la distribución de x y son los que se muestran en la tabla 12.3.

Tabla 12.3

x	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

En la tabla 12.4 se presentan las frecuencias observadas y las frecuencias esperadas.

Tabla 12.4

Observada	15	35	49	58	65	76	72	60	35	29	6
Esperada	13.9	27.8	41.7	55.6	69.5	83.4	69.5	55.6	41.7	27.8	13.9

Si en las celdas B1:L2 de una hoja de cálculo de EXCEL se introducen las frecuencias observadas y las frecuencias esperadas, en la celda B4 se introduce la expresión =(B1-B2)^2/B2, se hace clic y se arrastra desde B4 hasta L4 y las cantidades en B4:L4 se suman, se obtiene 10.34 como el valor de $\chi^2 = \sum_j ((o_j - e_j)^2/e_j)$.

El valor p que corresponde a 10.34 se obtiene mediante la expresión de EXCEL =CHIDIST(10.34,10). Este valor p es 0.411 y dado que es grande, no hay razón para pensar que el dado esté cargado.

TABLAS DE CONTINGENCIA

A tablas como la 12.1 en las que las frecuencias observadas ocupan un solo renglón se les llama *tablas de clasificación en un solo sentido*. Como el número de columnas es k, se les llama también *tablas* $1 \times k$ (que se lee "1 por k"). Por extensión de estas ideas, se obtienen *tablas de clasificación en dos sentidos*, o *tablas* $h \times k$, en las que las frecuencias observadas ocupan h renglones y k columnas. A estas tablas se les suele llamar *tablas de contingencia*.

En una tabla de contingencia $h \times k$, para cada frecuencia observada hay una *frecuencia esperada* (o *teórica*), que se calcula basándose en alguna hipótesis y sujetándose a las reglas de probabilidad. A las frecuencias que ocupan las *celdas* de una tabla de contingencia se les llama *frecuencias de celda*. Al total de las frecuencias de un renglón o de una columna se le llama *frecuencia marginal*.

Para investigar el grado de coincidencia entre las frecuencias observadas y las frecuencias esperadas se calcula el estadístico

$$\chi^2 = \sum_j \frac{(o_j - e_j)^2}{e_j} \tag{5}$$

donde la suma se realiza sobre todas las celdas de la tabla de contingencia y donde los símbolos o_j y e_j representan frecuencias, observada y esperada, en la celda j. Esta suma, que es análoga a la de la ecuación (*1*), contiene hk términos. La suma de todas las frecuencias observadas, que se denota N, es igual a la suma de todas las frecuencias esperadas [ver la ecuación (*2*)].

Como antes, el estadístico (*5*) tiene una distribución muestral que está dada, con una aproximación muy buena, por (*4*), siempre y cuando las frecuencias esperadas no sean demasiado pequeñas. El número de grados de libertad, ν, de esta distribución ji cuadrada es, para $h > 1$ y $k > 1$,

1. $\nu = (h - 1)(k - 1)$ si las frecuencias esperadas pueden calcularse sin necesidad de estimar parámetros poblacionales mediante estadísticos muestrales. Una demostración de esto se da en el problema 12.18.
2. $\nu = (h - 1)(k - 1) - m$ si las frecuencias esperadas sólo pueden calcularse estimando m parámetros poblacionales mediante estadísticos muestrales.

Las pruebas de significancia para tablas $h \times k$ son similares a las pruebas de significancia para tablas $1 \times k$. Las frecuencias esperadas se establecen basándose en la hipótesis H_0 de que se trate; una de las hipótesis más empleadas es que las dos clasificaciones son independientes una de otra.

Las tablas de contingencia pueden extenderse a dimensiones mayores. Así, se pueden tener, por ejemplo, tablas $h \times k \times 1$, en las que hay tres clasificaciones.

EJEMPLO 2 En la tabla 12.5 se presenta la manera en que las personas hacen sus declaraciones de impuestos y su nivel de estudios. La hipótesis nula es que la manera en que las personas hacen sus declaraciones de impuestos (usando software o sólo lápiz y papel) es independiente de su nivel de estudios. La tabla 12.5 es una tabla de contingencia.

Tabla 12.5

	Nivel de estudios		
Manera	Preparatoria	Licenciatura	Maestría
Computadora	23	35	42
Papel y lápiz	45	30	25

Empleando MINITAB para analizar estos datos se obtienen los resultados siguientes.

Prueba ji cuadrada: preparatoria, licenciatura, maestría

```
Los resultados esperados se muestran debajo de los observados
Las contribuciones de ji cuadrada se muestran debajo de los esperados

      preparatoria  licenciatura     maestría     Total
1               23            35           42       100
             34.00         32.50        33.50
             3.559         0.192        2.157

2               45            30           25       100
             34.00         32.50        33.50
             3.559         0.192        2.157

Total           68            65           67       200
Ji-Sq = 11.816, DF = 2, P-Value = 0.003
```

Debido a que el valor *p* es pequeño, se rechaza la hipótesis de independencia y se concluye que la manera en que se hace la declaración de impuestos y el nivel de educación no son independientes.

CORRECCIÓN DE YATES POR CONTINUIDAD

Cuando a datos discretos se aplican fórmulas para datos continuos, como se ha visto en capítulos anteriores, es necesario hacer una corrección por continuidad. Para el empleo de la distribución ji cuadrada hay una corrección similar. Esta corrección consiste en reescribir la ecuación (*1*) de la manera siguiente:

$$\chi^2 \text{ (corregida)} = \frac{(|o_1 - e_1| - 0.5)^2}{e_1} + \frac{(|o_2 - e_2| - 0.5)^2}{e_2} + \cdots + \frac{(|o_k - e_k| - 0.5)^2}{e_k} \tag{6}$$

y se le conoce como *corrección de Yates*. Para la ecuación (*5*) existe una modificación análoga.

En general, esta corrección sólo se hace cuando el número de grados de libertad es $\nu = 1$. Cuando se tienen muestras grandes, se obtiene prácticamente el mismo resultado que con χ^2 no corregida, pero cerca de los valores críticos pueden surgir dificultades (ver el problema 12.8). Cuando se tienen muestras pequeñas, donde cada una de las frecuencias esperadas está entre 5 y 10, quizá sea mejor comparar ambos valores de χ^2, el corregido y el no corregido. Si ambos valores conducen a la misma conclusión respecto a la hipótesis, por ejemplo al rechazo al nivel 0.05, es raro que se encuentren dificultades. Si ambos valores conducen a conclusiones diferentes se puede recurrir a aumentar el tamaño de la muestra, o si esto no es posible se pueden usar métodos de probabilidad en los que se emplee la *distribución multinomial* del capítulo 6.

FÓRMULAS SENCILLAS PARA CALCULAR χ^2

Para calcular χ^2 pueden deducirse fórmulas sencillas en las que únicamente se emplean las frecuencias esperadas. A continuación se dan las fórmulas para tablas de contingencia 2×2 y 2×3 (ver las tablas 12.6 y 12.7, respectivamente).

Tablas 2 × 2

$$\chi^2 = \frac{N(a_1 b_2 - a_2 b_1)^2}{(a_1 + b_1)(a_2 + b_2)(a_1 + a_2)(b_1 + b_2)} = \frac{N\Delta^2}{N_1 N_2 N_A N_B} \tag{7}$$

Tabla 12.6

	I	II	Total
A	a_1	a_2	N_A
B	b_1	b_2	N_B
Total	N_1	N_2	N

Tabla 12.7

	I	II	III	Total
A	a_1	a_2	a_3	N_A
B	b_1	b_2	b_3	N_B
Total	N_1	N_2	N_3	N

donde $\Delta = a_1b_2 - a_2b_1$, $N = a_1 + a_2 + b_1 + b_2$, $N_1 = a_1 + b_1$, $N_2 = a_2 + b_2$, $+ b_2$, $N_A = a_1 + a_2$ y $N_B = b_1 + b_2$ (ver problema 12.19). Empleando la corrección de Yates, esta ecuación se convierte en

$$\chi^2(\text{corregida}) = \frac{N(|a_1b_2 - a_2b_1| - \frac{1}{2}N)^2}{(a_1 + b_1)(a_2 + b_2)(a_1 + a_2)(b_1 + b_2)} = \frac{N(|\Delta| - \frac{1}{2}N)^2}{N_1N_2N_AN_B} \tag{8}$$

Tablas 2 × 3

$$\chi^2 = \frac{N}{N_A}\left[\frac{a_1^2}{N_1} + \frac{a_2^2}{N_2} + \frac{a_3^2}{N_3}\right] + \frac{N}{N_B}\left[\frac{b_1^2}{N_1} + \frac{b_2^2}{N_2} + \frac{b_3^2}{N_3}\right] - N \tag{9}$$

donde se ha empleado el resultado general válido para todas las tablas de contingencia (ver el problema 12.43):

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N \tag{10}$$

La fórmula (*9*) puede generalizarse a tablas $2 \times k$ donde $k > 3$ (ver el problema 12.46).

COEFICIENTE DE CONTINGENCIA

Una medida del grado de relación, asociación o dependencia entre las clasificaciones en una tabla de contingencia es la dada por

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \tag{11}$$

que se conoce como *coeficiente de contingencia*. Cuanto mayor sea el valor de C mayor será el grado de relación entre las clasificaciones. El valor máximo de C está determinado por la cantidad de renglones y columnas de la tabla de contingencia y este valor nunca es mayor a 1. Si k es la cantidad de renglones y columnas en una tabla de contingencia, el valor máximo de C es $\sqrt{(k-1)/k}$ (ver los problemas 12.22, 12.52 y 12.53).

EJEMPLO 3 Encontrar el coeficiente de contingencia correspondiente al ejemplo 2.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{11.816}{11.816 + 200}} = 0.236$$

CORRELACIÓN DE ATRIBUTOS

Como las clasificaciones de una tabla de contingencia suelen describir características de personas u objetos, a estas clasificaciones se les suele llamar *atributos* y a su grado de dependencia, asociación o relación se le llama *correlación de atributos*. Para tablas $k \times k$, el coeficiente de correlación entre atributos (o clasificaciones) se define como

$$r = \sqrt{\frac{\chi^2}{N(k-1)}} \tag{12}$$

este coeficiente se encuentra entre 0 y 1 (ver el problema 12.24). En tablas 2×2 en las que $k = 2$, a la correlación se le conoce como *correlación tetracórica*.

En el capítulo 14 se considera el problema general de la correlación entre variables numéricas.

PROPIEDAD ADITIVA DE χ^2

Supóngase que como resultado de la repetición de un experimento se obtienen los valores muestrales de χ^2 dados por $\chi_1^2, \chi_2^2, \chi_3^2, \ldots$ con $\nu_1, \nu_2, \nu_3, \ldots$ grados de libertad, respectivamente. Entonces el resultado de todos estos experimentos puede considerarse equivalente al valor χ^2 dado por $\chi_1^2 + \chi_2^2 + \chi_3^2 + \cdots$ con $\nu_1 + \nu_2 + \nu_3 + \cdots$ grados de libertad (ver el problema 12.25).

PROBLEMAS RESUELTOS

LA PRUEBA JI CUADRADA

12.1 En 200 lanzamientos de una moneda se obtienen 115 caras y 85 cruces. Pruebe la hipótesis de que la moneda no está cargada a los niveles de significancia: *a*) 0.05 y *b*) 0.01, empleando el apéndice IV y *c*) pruebe esta hipótesis calculando el valor *p* y comparándolo con los niveles 0.05 y 0.01.

SOLUCIÓN

Las frecuencias observadas de caras y cruces son $o_1 = 115$ y $o_2 = 85$, respectivamente, y las frecuencias esperadas de caras y cruces (si la moneda no está cargada) son $e_1 = 100$ y $e_2 = 100$, respectivamente. Por lo tanto,

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} = \frac{(115 - 100)^2}{100} + \frac{(85 - 100)^2}{100} = 4.50$$

Dado que el número de categorías, o clases (caras, cruces), es $k = 2$, $\nu = k - 1 = 2 - 1 = 1$.

a) El valor crítico χ_{95}^2 para 1 grado de libertad es 3.84. Por lo tanto, como $4.50 > 3.84$, al nivel de significancia 0.05 se rechaza la hipótesis de que la moneda no está cargada.

b) El valor crítico χ_{99}^2 para 1 grado de libertad es 6.63. Por lo tanto, como $4.50 < 6.63$, al nivel de significancia 0.01 no se puede rechazar la hipótesis de que la moneda no está cargada.

Se concluye que los resultados encontrados *tal vez sean significativos* y que la moneda *quizás esté cargada*. Para comparar este método con métodos usados antes, ver el problema 12.3.

Empleando EXCEL, el valor *p* se obtiene mediante =CHIDIST(4.5,1) que da como resultado 0.0339. Y empleando el método del valor *p* se ve que los resultados son significativos a 0.05, pero no a 0.01. Cualquiera de estos métodos puede emplearse para realizar la prueba.

12.2 Se repite el problema 12.1 empleando la corrección de Yates.

SOLUCIÓN

$$\chi^2 \text{ (corregida)} = \frac{(|o_1 - e_1| - 0.5)^2}{e_1} + \frac{(|o_2 - e_2| - 0.5)^2}{e_2} = \frac{(|115 - 100| - 0.5)^2}{100} + \frac{(|85 - 100| - 0.5)^2}{100}$$

$$= \frac{(14.5)^2}{100} + \frac{(14.5)^2}{100} = 4.205$$

Como $4.205 > 3.84$ y $4.205 < 6.63$, las conclusiones a las que se llegó en el problema 12.1 son válidas. Para hacer una comparación con los métodos anteriores, ver el problema 12.3.

12.3 Resolver el problema 12.1 empleando la aproximación normal a la distribución binomial.

SOLUCIÓN

De acuerdo con la hipótesis de que la moneda no está cargada, la media y la desviación estándar de la cantidad de caras esperadas en 200 lanzamientos de una moneda son $\mu = Np = (200)(0.5) = 100$ y $\sigma = \sqrt{Npq} = \sqrt{(200)(0.5)(0.5)} = 7.07$, respectivamente.

Primer método

$$115 \text{ caras en unidades estándar} = \frac{115 - 100}{7.07} = 2.12$$

Al nivel de significancia 0.05, empleando una prueba de dos colas, la hipótesis de que la moneda no está cargada se rechaza si la puntuación z que se obtenga cae fuera del intervalo -1.96 a 1.96. Al nivel 0.01 el intervalo correspondiente es -2.58 a 2.58. Se concluye (como en el problema 12.1) que la hipótesis puede rechazarse al nivel 0.05, pero no al nivel 0.01.

Obsérvese que el cuadrado de la puntuación estándar anterior es $(2.12)^2 = 4.50$, que es igual al valor de χ^2 obtenido en el problema 12.1. Éste es siempre el caso en una prueba ji cuadrada con dos categorías (ver problema 12.10).

Segundo método

Usando la corrección por continuidad 115 o más caras es equivalente a 114.5 o más caras. Entonces 114.5 en unidades estándar $= (114.5 - 100)/7.07 = 2.05$. Esto conduce a la misma conclusión obtenida con el primer método.

Obsérvese que el cuadrado de la puntuación estándar es $(2.05)^2 = 4.20$, valor que coincide con el valor de χ^2 corregido por continuidad empleando la corrección de Yates en el problema 12.2. Éste es siempre el caso en una prueba ji cuadrada en la que haya dos categorías y se emplee la corrección de Yates.

12.4 En la tabla 12.8 se muestran las frecuencias observadas y las frecuencias esperadas al lanzar un dado 120 veces.

a) Pruebe la hipótesis de que el dado no está cargado calculando χ^2 y comparando el estadístico de prueba encontrado con el valor crítico correspondiente al nivel de significancia 0.05.

b) Calcule el valor p y compárelo con 0.05 para probar la hipótesis.

Tabla 12.8

Cara del dado	1	2	3	4	5	6
Frecuencias observadas	25	17	15	23	24	16
Frecuencias esperadas	20	20	20	20	20	20

SOLUCIÓN

$$\chi^2 = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \frac{(o_3 - e_3)^2}{e_3} + \frac{(o_4 - e_4)^2}{e_4} + \frac{(o_5 - e_5)^2}{e_5} + \frac{(o_6 - e_6)^2}{e_6}$$

$$= \frac{(25 - 20)^2}{20} + \frac{(17 - 20)^2}{20} + \frac{(15 - 20)^2}{20} + \frac{(23 - 20)^2}{20} + \frac{(24 - 20)^2}{20} + \frac{(16 - 20)^2}{20} = 5.00$$

a) Empleando EXCEL, el valor crítico correspondiente a 0.05 se obtiene mediante la expresión =CHIINV(0.05,5), que da 11.0705. El valor encontrado para el estadístico de prueba es 5.00. Como el valor encontrado para el estadístico de prueba no está en la región crítica 0.05, no se rechaza la hipótesis nula de que el dado no esté cargado.

b) Empleando EXCEL, el valor p se obtiene mediante la expresión =CHIDIST(5.00,5), que da 0.4159. Como el valor p no es menor a 0.05, no se rechaza la hipótesis nula de que el dado no esté cargado.

12.5 En la tabla 12.9 se muestra la distribución de los dígitos 0, 1, 2, ..., 9 en los 250 dígitos de una tabla de números aleatorios. *a*) Encontrar el valor del estadístico de prueba χ^2, *b*) encontrar el valor crítico correspondiente a $\alpha = 0.01$ y dar una conclusión y *c*) encontrar el valor *p* correspondiente al valor encontrado en el inciso *a*) y dar una conclusión para $\alpha = 0.01$.

Tabla 12.9

Dígito	0	1	2	3	4	5	6	7	8	9
Frecuencias observadas	17	31	29	18	14	20	35	30	20	36
Frecuencias esperadas	25	25	25	25	25	25	25	25	25	25

SOLUCIÓN

a) $$\chi^2 = \frac{(17-25)^2}{25} + \frac{(31-25)^2}{25} + \frac{(29-25)^2}{25} + \frac{(18-25)^2}{25} + \cdots + \frac{(36-25)^2}{25} = 23.3$$

b) El valor crítico correspondiente a 0.01 se obtiene mediante la expresión =CHIINV(0.01,9) y es 21.6660. Como el valor obtenido para χ^2 es mayor a este valor, se rechaza la hipótesis de que estos números sean aleatorios.

c) Empleando EXCEL, el valor *p* se obtiene mediante la expresión =CHIDIST(23.3,9) y es 0.0056, que es menor a 0.01. De manera que con la técnica del valor *p* se rechaza la hipótesis nula.

12.6 En un experimento empleando chícharos, Gregor Mendel observó que 315 eran redondos y amarillos, 108 eran redondos y verdes, 101 eran deformes y amarillos, y 32 eran deformes y verdes. De acuerdo con su teoría sobre la herencia, estas cantidades debían estar en la proporción 9:3:3:1. ¿Existe alguna evidencia que haga dudar de su teoría a los niveles de significancia: *a*) 0.01 y *b*) 0.05?

SOLUCIÓN

La cantidad total de chícharos es $315 + 108 + 101 + 35 = 556$. Como las cantidades esperadas están en la proporción 9:3:3:1 (y $9 + 3 + 3 + 1 = 16$), se esperaría que hubiera

$$\tfrac{9}{16}(556) = 312.75 \text{ redondos y amarillos} \qquad \tfrac{3}{16}(556) = 104.25 \text{ deformes y amarillos}$$
$$\tfrac{3}{16}(556) = 104.25 \text{ redondos y verdes} \qquad \tfrac{1}{16}(556) = 34.75 \text{ deformes y verdes}$$

Por lo tanto, $$\chi^2 = \frac{(315-312.75)^2}{312.75} + \frac{(108-104.25)^2}{104.25} + \frac{(101-104.25)^2}{104.25} + \frac{(32-34.75)^2}{34.75} = 0.470$$

Dado que hay cuatro categorías, $k = 4$ y el número de grados de libertad es $\nu = 4 - 1 = 3$.

a) Para $\nu = 3$, $\chi^2_{.99} = 11.3$; por lo tanto, al nivel 0.01 no puede rechazarse su teoría.

b) Para $\nu = 3$, $\chi^2_{.95} = 7.81$; por lo tanto, al nivel 0.05 no puede rechazarse su teoría.

Se concluye que sí hay coincidencia entre la teoría y la experimentación.

Obsérvese que para 3 grados de libertad $\chi^2_{.05} = 0.352$ y $\chi^2 = 0.470 > 0.352$. Por lo tanto, aunque la coincidencia sea buena, el resultado obtenido está sujeto a una cantidad razonable de error muestral.

12.7 En una urna hay una cantidad grande de canicas de cuatro colores: rojas, anaranjadas, amarillas y verdes. En una muestra de 12 canicas, tomada de la urna en forma aleatoria, se encuentran 2 canicas rojas, 4 canicas anaranjadas, 4 canicas amarillas y 1 canica verde. Probar la hipótesis de que en la urna las canicas de los distintos colores están en la misma proporción.

SOLUCIÓN

Bajo la hipótesis de que en la urna hay la misma proporción de canicas de cada color, se esperaría que en una muestra de 12 canicas hubiera 3 de cada color. Como las cantidades esperadas son menores a 5, la aproximación ji cuadrada será errónea. Para evitar esto se fusionan categorías de manera que el tamaño de cada categoría sea por lo menos 5.

Si se desea rechazar la hipótesis habrá que combinar las categorías de manera que la evidencia contra la hipótesis sea la mejor posible. En tal caso, esto se logra formando las categorías "rojas o verdes" y "anaranjadas o amarillas", con lo cual las muestras serán de 3 y 9 canicas, respectivamente. Como la cantidad esperada en cada categoría, de acuerdo con la hipótesis de proporciones iguales, es 6, se tiene

$$\chi^2 = \frac{(3-6)^2}{6} + \frac{(9-6)^2}{6} = 3$$

Para $\nu = 2 - 1 = 1$, $\chi^2_{.95} = 3.84$. Por lo tanto, al nivel de significancia 0.05 no se puede rechazar la hipótesis (aunque sí al nivel de significancia 0.10). Por supuesto que los resultados obtenidos pueden deberse únicamente a la casualidad aun cuando los distintos colores estén en la misma proporción.

Otro método

Empleando la corrección de Yates, se encuentra

$$\chi^2 = \frac{(|3-6|-0.5)^2}{6} + \frac{(|9-6|-0.5)^2}{6} = \frac{(2.5)^2}{6} + \frac{(2.5)^2}{6} = 2.1$$

lo que conduce a la misma conclusión obtenida antes. Esto era de esperarse, ya que la corrección de Yates siempre *reduce* el valor de χ^2.

Nótese que empleando la aproximación χ^2, aun cuando las frecuencias son demasiado pequeñas, se obtiene

$$\chi^2 = \frac{(2-3)^2}{3} + \frac{(5-3)^2}{3} + \frac{(4-3)^2}{3} + \frac{(1-3)^2}{3} = 3.33$$

Como $\nu = 4 - 1 = 3$, $\chi^2_{.95} = 7.81$ y se llega a la misma conclusión que antes. Infortunadamente, cuando las frecuencias son pequeñas, la aproximación χ^2 es pobre; por lo tanto, cuando no sea recomendable combinar frecuencias hay que recurrir a los métodos exactos de probabilidad del capítulo 6.

12.8 En 360 lanzamientos de un par de dados se obtuvo 74 veces un 7 y 24 veces un 11. Empleando como nivel de significancia 0.05 pruebe la hipótesis de que el dado no está cargado.

SOLUCIÓN

Un par de dados pueden caer de 36 maneras. El número once se puede obtener de 6 maneras y el número siete de 2 maneras. Entonces $\Pr\{\text{siete}\} = \frac{6}{36} = \frac{1}{6}$ y $\Pr\{\text{once}\} = \frac{2}{36} = \frac{1}{18}$. Por lo tanto, en 360 lanzamientos se esperan $\frac{1}{6}(360) = 60$ sietes y $\frac{1}{18}(360) = 20$ onces, de manera que

$$\chi^2 = \frac{(74-60)^2}{60} + \frac{(24-20)^2}{20} = 4.07$$

Para $\nu = 2 - 1 = 1$, $\chi^2_{.95} = 3.84$. Por lo tanto, como $4.07 > 3.84$, se estará inclinado a rechazar la hipótesis de que el dado no está cargado. Sin embargo, empleando la corrección de Yates se encuentra

$$\chi^2 \text{ (corregida)} = \frac{(|74-60|-0.5)^2}{60} + \frac{(|24-20|-0.5)^2}{20} = \frac{(13.5)^2}{60} + \frac{(3.5)^2}{20} = 3.65$$

Así, de acuerdo con el valor de χ^2 corregida, no se puede rechazar la hipótesis al nivel 0.05.

En general, con muestras grandes como las que se tienen en este caso, los resultados empleando la corrección de Yates son más confiables que sin usar la corrección de Yates. Sin embargo, como aun el valor corregido de χ^2 está tan cercano al valor crítico, se estará indeciso para tomar una decisión en un sentido o en otro. En tales casos, quizá lo mejor sea aumentar el tamaño de la muestra si, por alguna razón, se está especialmente interesado en el nivel 0.05; si no es así, se puede rechazar la hipótesis a algún otro nivel (por ejemplo, al nivel 0.10) si esto es satisfactorio.

12.9 Se estudian 320 familias de 5 hijos cada una y se encuentra la distribución que se muestra en la tabla 12.10. ¿Este resultado es consistente con la hipótesis de que el nacimiento de un hombre o de una mujer es igualmente probable?

Tabla 12.10

Cantidad de niños y niñas	5 niños 0 niñas	4 niños 1 niña	3 niños 2 niñas	2 niños 3 niñas	1 niño 4 niñas	0 niños 5 niñas	Total
Cantidad de familias	18	56	110	88	40	8	320

SOLUCIÓN

Sea $p =$ probabilidad de que nazca un hombre y $q = 1 - p =$ probabilidad de que nazca una mujer. Entonces las probabilidades de (5 niños), (4 niños y 1 niña), ..., (5 niñas) están dadas por los términos de la expansión binomial

$$(p+q)^5 = p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5$$

Si $p = q = \frac{1}{2}$, se tiene

$$\Pr\{5 \text{ niños y } 0 \text{ niñas}\} = (\tfrac{1}{2})^5 = \tfrac{1}{32} \qquad \Pr\{2 \text{ niños y } 3 \text{ niñas}\} = 10(\tfrac{1}{2})^2(\tfrac{1}{2})^3 = \tfrac{10}{32}$$

$$\Pr\{4 \text{ niños y } 1 \text{ niña}\} = 5(\tfrac{1}{2})^4(\tfrac{1}{2}) = \tfrac{5}{32} \qquad \Pr\{1 \text{ niño y } 4 \text{ niñas}\} = 5(\tfrac{1}{2})(\tfrac{1}{2})^4 = \tfrac{5}{32}$$

$$\Pr\{3 \text{ niños y } 2 \text{ niñas}\} = 10(\tfrac{1}{2})^3(\tfrac{1}{2})^2 = \tfrac{10}{32} \qquad \Pr\{0 \text{ niños y } 5 \text{ niñas}\} = (\tfrac{1}{2})^5 = \tfrac{1}{32}$$

Por lo que las cantidades esperadas de familias con 5, 4, 3, 2, 1 y 0 niños se obtienen multiplicando las probabilidades anteriores por 320, y los resultados son 10, 50, 100, 100, 50 y 10, respectivamente. Por lo tanto,

$$\chi^2 = \frac{(18-10)^2}{10} + \frac{(56-50)^2}{50} + \frac{(110-100)^2}{100} + \frac{(88-100)^2}{100} + \frac{(40-50)^2}{50} + \frac{(8-10)^2}{10} = 12.0$$

Como $\chi^2_{.95} = 11.1$ y $\chi^2_{.99} = 15.1$ para $\nu = 6 - 1 = 5$ grados de libertad, la hipótesis nula puede rechazarse al nivel de significancia 0.05, pero no al nivel de significancia 0.01. De manera que se concluye que los resultados tal vez sean significativos y que el nacimiento de hombres y mujeres no es igualmente probable.

12.10 En 500 personas estudiadas se encontró que la semana pasada 155 de ellas habían rentado por lo menos un video. Empleando una prueba de dos colas y $\alpha = 0.05$, probar la hipótesis de que la semana pasada el 25% de la población rentó por lo menos un video. Realizar la prueba empleando tanto la distribución normal estándar como la distribución ji cuadrada. Mostrar que la prueba ji cuadrada con sólo dos categorías es equivalente a la prueba de significancia para proporciones dada en el capítulo 10.

SOLUCIÓN

Si la hipótesis nula es verdadera, entonces $\mu = Np = 500(0.25) = 125$ y $\sigma = \sqrt{Npq} = \sqrt{500(0.25)(0.75)} = 9.68$. El estadístico de prueba calculado es $Z = (155 - 125)/9.68 = 3.10$. Los valores críticos son ± 1.96, por lo que la hipótesis nula se rechaza.

La solución empleando la distribución ji cuadrada se halla empleando los resultados que se muestran en la tabla 12.11.

Tabla 12.11

Frecuencias	Rentaron videos	No rentaron videos	Total
Observadas	155	345	500
Esperadas	125	375	500

El estadístico ji cuadrada calculado se obtiene como sigue:

$$\chi^2 = \frac{(155-125)^2}{125} + \frac{(345-375)^2}{375} = 9.6$$

El valor crítico para un grado de libertad es 3.84, por lo que se rechazó la hipótesis nula. Obsérvese que $(3.10)^2 = 9.6$ y que $(\pm 1.96)^2 = 3.84$, o sea $Z^2 = \chi^2$. Los dos procedimientos son equivalentes.

12.11 *a*) Probar que la fórmula (*1*) de este capítulo puede escribirse como

$$\chi^2 = \sum \frac{o_j^2}{e_j} - N$$

b) Utilizar el resultado de *a*) para comprobar el valor de χ^2 calculado en el problema 12.6.

SOLUCIÓN

a) Por definición

$$\chi^2 = \sum \frac{(o_j - e_j)^2}{e_j} = \sum \left(\frac{o_j^2 - 2o_j e_j + e_j^2}{e_j} \right)$$

$$= \sum \frac{o_j^2}{e_j} - 2 \sum o_j + \sum e_j = \sum \frac{o_j^2}{e_j} - 2N + N = \sum \frac{o_j^2}{e_j} - N$$

donde se ha empleando la fórmula (*2*) de este capítulo.

b) $$\chi^2 = \sum \frac{o_j^2}{e_j} - N = \frac{(315)^2}{312.75} + \frac{(108)^2}{104.25} + \frac{(101)^2}{104.25} + \frac{(32)^2}{34.75} - 556 = 0.470$$

BONDAD DE AJUSTE

12.12 Un jugador de tenis se entrena jugando series de tres juegos; lleva un registro de los juegos perdidos y ganados en estas series a lo largo del año. Su registro muestra que de 250 días, 25 días ganó 0 juegos, 75 días ganó 1 juego, 125 días ganó 2 juegos y 25 días ganó 3 juegos. Con $\alpha = 0.05$, probar que $X =$ cantidad de juegos ganados, en las series de 3, está distribuida en forma binomial.

SOLUCIÓN

La cantidad media de juegos ganados en estas series de 3 juegos es $(0 \times 25 + 1 \times 75 + 2 \times 125 + 3 \times 25)/250 = 1.6$. Si X es binomial, la media es $np = 3p$, lo cual igualándolo al estadístico 1.6 y despejando p permite encontrar que $p = 0.53$. Se desea probar que X es binomial con $n = 3$ y $p = 0.53$. Si X es binomial con $p = 0.53$, su distribución y el número esperado de juegos ganados son los que muestran los siguientes resultados de EXCEL. Obsérvese que las probabilidades binomiales $p(x)$ se encontraron ingresando =BINOMDIST(A2,3,0.53,0) y haciendo clic y arrastrando desde B2 hasta B5. De esta manera se obtuvieron los valores que se muestran a continuación bajo $p(x)$.

x	p(x)	Ganados esperados	Ganados observados
0	0.103823	25.95575	25
1	0.351231	87.80775	75
2	0.396069	99.01725	125
3	0.148877	37.21925	25

La cantidad de juegos ganados esperados se encuentra multiplicando los valores de $p(x)$ por 250.

$$\chi^2 = \frac{(25 - 30.0)^2}{30.0} + \frac{(75 - 87.8)^2}{87.8} + \frac{(125 - 99.0)^2}{99.0} + \frac{(25 - 37.2)^2}{37.2} = 12.73.$$

Como la cantidad de parámetros necesarios para estimar las frecuencias esperadas es $m = 1$ (a saber, el parámetro p de la distribución binomial), $v = k - 1 - m = 4 - 1 - 1 = 2$. El valor p se obtiene mediante la expresión de EXCEL =CHIDIST(12.73,2) = 0.0017, por lo que se rechaza la hipótesis de que la variable X esté distribuida en forma binomial.

12.13 El número de horas por semana que 200 estudiantes universitarios usan Internet se ha agrupado en las clases 0 a 3, 4 a 7, 8 a 11, 12 a 15, 16 a 19, 20 a 23 y 24 a 27, cuyas frecuencias observadas son 12, 25, 36, 45, 34, 31 y 17. A partir de estos datos se obtiene la media y la desviación estándar de estos datos agrupados. La hipótesis

nula es que estos datos están distribuidos normalmente. De acuerdo con la media y con la desviación estándar encontradas y suponiendo que la distribución sea normal, se obtienen las frecuencias esperadas que, redondeadas, son las siguientes: 10, 30, 40, 50, 36, 28 y 6.

a) Encontrar χ^2.
b) ¿Cuántos grados de libertad tiene χ^2?
c) Empleando EXCEL, encontrar el valor crítico del 5% y dar las conclusiones al 5%.
d) Empleando EXCEL, hallar el valor *p* para el resultado.

SOLUCIÓN

a) En la figura 12-1 se muestra parte de la hoja de cálculo de EXCEL. En C2 se ingresa =(A2-B2)^2/B2, se hace clic y se arrastra desde C2 hasta C8. En C9 se ingresa =SUM(C2:C8). Como se ve, $\chi^2 = 22.7325$.

	A	B	C	D
1	observed	expected	(O - E)^2/E	
2	12	10	0.4	
3	25	30	0.8333333	
4	36	40	0.4	
5	45	50	0.5	
6	34	36	0.1111111	
7	31	28	0.3214286	
8	17	6	20.166667	
9			22.73254	
10				

Figura 12-1 EXCEL, parte de la hoja de cálculo para el problema 12.13.

b) Como el número de parámetros empleados para estimar las frecuencias esperadas es $m = 2$ (comúnmente son la media μ y la desviación estándar σ de una distribución normal), $\nu = k - 1 - m = 7 - 1 - 2 = 4$. Obsérvese que no es necesario combinar clases, ya que todas las frecuencias esperadas son mayores a 5.
c) El valor crítico de 5% se obtiene mediante =CHIINV(0.05,4) y es 9.4877. Como 22.73 es mayor al valor crítico, se rechaza la hipótesis nula de que los datos provengan de una distribución normal.
d) El valor *p* se encuentra mediante =CHIDIST(22.7325,4), que da valor $p = 0.000143$.

TABLAS DE CONTINGENCIA

12.14 Se repite el problema 10.20 usando, primero, la prueba ji cuadrada, y después MINITAB. Comparar las dos soluciones.

SOLUCIÓN

En la tabla 12.12*a*) se presentan las condiciones del problema. Bajo la hipótesis nula de que el suero no tiene efecto alguno, se esperaría que en cada grupo se recuperaran 70 personas y 30 no, como se muestra en la tabla 12.12*b*). Obsérvese que la hipótesis nula es equivalente a afirmar que la recuperación es *independiente* del uso del suero (es decir, que las clasificaciones son independientes).

Tabla 12.12*a*) Frecuencias observadas

	Recuperados	No recuperados	Total
Grupo A (usan el suero)	75	25	100
Grupo B (no usan el suero)	65	35	100
Total	140	60	200

Tabla 12.12*b*) Frecuencias esperadas bajo H_0

	Recuperados	No recuperados	Total
Grupo A (usan el suero)	70	30	100
Grupo B (no usan el suero)	70	30	100
Total	140	60	200

$$\chi^2 = \frac{(75-70)^2}{70} + \frac{(65-70)^2}{70} + \frac{(25-30)^2}{30} + \frac{(35-30)^2}{30} = 2.38$$

Para determinar el número de grados de libertad, considérese la tabla 12.13, que es la misma tabla 12.12, excepto que sólo muestra los totales. Es claro que en cualquiera de las cuatro celdas vacías sólo se tiene la libertad de colocar un número, ya que una vez hecho esto los números de las celdas restantes quedan determinados de manera única por los totales dados. Por lo tanto, hay 1 grado de libertad.

Tabla 12.13

	Recuperados	No recuperados	Total
Grupo A			100
Grupo B			100
Total	140	60	200

Otro método

Empleando la fórmula (ver problema 12.18), $\nu = (h-1)(k-1) = (2-1)(2-1) = 1$. Como $\chi^2_{.95} = 3.84$ para 1 grado de libertad y como $\chi^2 = 2.38 < 3.84$, se concluye que los resultados *no son significativos* al nivel 0.05. Por lo tanto, no se puede rechazar H_0 a este nivel, y se concluye que el suero no es efectivo o se aplaza la decisión hasta tener más resultados.

Obsérvese que $\chi^2 = 2.38$ es el cuadrado de la puntuación z, $z = 1.54$, que se obtuvo en el problema 10.20. En general, la prueba ji cuadrada para proporciones muestrales en una tabla de contingencia 2×2 es equivalente a una prueba de significancia para la diferencia entre proporciones usando la aproximación normal.

Nótese también que una prueba de una cola empleando χ^2 es equivalente a una prueba de dos colas empleando χ, ya que, por ejemplo, $\chi^2 > \chi^2_{.95}$ corresponde a $\chi > \chi_{.95}$ o $\chi < -\chi_{.95}$. Como en tablas 2×2 χ^2 es el cuadrado de la puntuación z, se sigue que, en este caso, χ es igual a z. Por lo tanto, el rechazo de una hipótesis al nivel 0.05 empleando χ^2 es equivalente al rechazo de una prueba de dos colas al nivel 0.10 empleando z.

Prueba ji cuadrada: recuperación, no recuperación

```
Los resultados esperados se muestran debajo de los observados
Las contribuciones de ji cuadrada se muestran debajo de los resultados esperados

       Recuperación      No recuperación      Total
1                75                   25        100
              70.00                30.00
              0.357                0.833
2                65                   35        100
              70.00                30.00
              0.357                0.833
Total           140                   60        200
Ji-Sq=2.381, DF=1, P-Value=0.123
```

12.15 Resolver el problema 12.14 empleando la corrección de Yates.

SOLUCIÓN

$$\chi^2(\text{corregida}) = \frac{(|75-70|-0.5)^2}{70} + \frac{(|65-70|-0.5)^2}{70} + \frac{(|25-30|-0.5)^2}{30} + \frac{(|35-30|-0.5)^2}{30} = 1.93$$

Por lo tanto, las conclusiones a las que se llegó en el problema 12.14 son correctas. Esto era de suponer sabiendo que la corrección de Yates siempre hace disminuir el valor de χ^2.

12.16 Una empresa de teléfonos celulares realiza una encuesta para determinar la proporción de personas que tienen teléfono celular en los distintos grupos de edad. En la tabla 12.14 se muestran los resultados obtenidos en 100 hogares. Probar la hipótesis de que en los diferentes grupos de edad, las proporciones de personas que tienen teléfono celular son las mismas.

Tabla 12.14

Teléfono celular	18-24	25-54	55-64	≥65	Total
Sí	50	80	70	50	250
No	200	170	180	200	750
Total	250	250	250	250	1 000

SOLUCIÓN

De acuerdo con la hipótesis de que la proporción de personas que tienen teléfono celular es la misma en los distintos grupos de edad, 250/1 000 = 25% es una estimación del porcentaje de personas que tienen teléfono celular en cada grupo de edad y 75% es una estimación del porcentaje de personas que no tienen teléfono celular en cada grupo de edad. En la tabla 12.15 se presentan las frecuencias esperadas en cada grupo de edad.

El valor del estadístico ji cuadrada se puede encontrar como se muestra en la tabla 12.16.

El número de grados de libertad para la distribución ji cuadrada es $\nu = (h-1)(k-1) = (2-1)(4-1) = 3$. Como $\chi^2_{.95} = 7.81$, y 14.3 es mayor que 7.81, se rechaza la hipótesis nula y se concluye que los porcentajes en los cuatro grupos de edad no son los mismos.

Tabla 12.15

Teléfono celular	18-24	25-54	55-64	≥65	Total
Sí	25% de 250 = 62.5	25% de 250 = 62.5	25% de 250 = 62.5	25% de 250 = 62.5	250
No	75% de 250 = 187.5	75% de 250 = 187.5	75% de 250 = 187.5	75% de 250 = 187.5	750
Total	250	250	250	250	1 000

Tabla 12.16

Renglón, columna	o	e	$(o-e)$	$(o-e)^2$	$(o-e)^2/e$
1, 1	50	62.5	−12.5	156.25	2.5
1, 2	80	62.5	17.5	306.25	4.9
1, 3	70	62.5	7.5	56.25	0.9
1, 4	50	62.5	−12.5	156.25	2.5
2, 1	200	187.5	12.5	156.25	0.8
2, 2	170	187.5	−17.5	306.25	1.6
2, 3	180	187.5	−7.5	56.25	0.3
2, 4	200	187.5	12.5	156.25	0.8
Suma	1 000	1 000	0		14.3

12.17 Utilizar MINITAB para resolver el problema 12.16.

SOLUCIÓN

A continuación se presenta la solución que da MINITAB al problema 12.16. Las cantidades observadas y esperadas se presentan junto con los cálculos del estadístico de prueba. Obsérvese que la hipótesis nula se rechazará a cualquier nivel de significancia mayor a 0.002.

Muestra de datos

```
Fila    18-24     25-54     55-64     65 o más
   1       50        80        70           50
   2      200       170       180          200

MTB > chisquare c1-c4
```

Prueba ji cuadrada

```
Los resultados esperados se muestran debajo de los observados

           18-24     25-54     55-64  65 o más     Total
      1       50        80        70        50       250
            62.50     62.50     62.50     62.50

      2      200       170       180       200       750
           187.50    187.50    187.50    187.50

  Total      250       250       250       250     1 000
Ji-Sq=     2.500 +   4.900 +   0.900 +   2.500 +
           0.833 +   1.633 +   0.300 +   0.833 = 14.400
DF = 3, P-Value = 0.002
```

12.18 Mostrar que en una tabla de contingencia de $h \times k$, el número de grados de libertad es $(h-1)(k-1)$, donde $h > 1$ y $k > 1$.

SOLUCIÓN

En una tabla con h renglones y k columnas únicamente se puede dejar sin introducir un número en cada renglón y en cada columna, ya que estos números se determinan conociendo los totales de cada columna y de cada renglón. Por lo tanto, sólo se tiene la libertad de colocar $(h-1)(k-1)$ números en la tabla, los números restantes quedan determinados automáticamente y de manera única. Por lo tanto, el número de grados de libertad es $(h-1)(k-1)$. Obsérvese que este resultado es válido si se conocen los parámetros poblacionales necesarios para obtener las frecuencias esperadas.

12.19 *a*) Demostrar que para la tabla de contingencia 2×2 que se muestra en la tabla 12.17*a*),

$$\chi^2 = \frac{N(a_1 b_2 - a_2 b_1)^2}{N_1 N_2 N_A N_B}$$

b) Ilustrar el resultado de *a*) empleando los datos del problema 12.14.

Tabla 12.17*a*) Resultados observados

	I	II	Total
A	a_1	a_2	N_A
B	b_1	b_2	N_B
Total	N_1	N_2	N

Tabla 12.17*b*) Resultados esperados

	I	II	Total
A	$N_1 N_A / N$	$N_2 N_A / N$	N_A
B	$N_1 N_B / N$	$N_2 N_B / N$	N_B
Total	N_1	N_2	N

SOLUCIÓN

a) Como en el problema 12.14, los resultados esperados, basándose en la hipótesis nula, se presentan en la tabla 12.17*b*). Entonces,

$$\chi^2 = \frac{(a_1 - N_1N_A/N)^2}{N_1N_A/N} + \frac{(a_2 - N_2N_A/N)^2}{N_2N_A/N} + \frac{(b_1 - N_1N_B/N)^2}{N_1N_B/N} + \frac{(b_2 - N_2N_B/N)^2}{N_2N_B/N}$$

Pero $$a_1 - \frac{N_1N_A}{N} = a_1 - \frac{(a_1 + b_1)(a_1 + a_2)}{a_1 + b_1 + a_2 + b_2} = \frac{a_1b_2 - a_2b_1}{N}$$

De manera similar $$a_2 - \frac{N_2N_A}{N} \quad \text{y} \quad b_1 - \frac{N_1N_B}{N} \quad \text{y} \quad b_2 - \frac{N_2N_B}{N}$$

son también igual a $$\frac{a_1b_2 - a_2b_1}{N}$$

Por lo tanto, se puede escribir

$$\chi^2 = \frac{N}{N_1N_A}\left(\frac{a_1b_2 - a_2b_1}{N}\right)^2 + \frac{N}{N_2N_A}\left(\frac{a_1b_2 - a_2b_1}{N}\right)^2 + \frac{N}{N_1N_B}\left(\frac{a_1b_2 - a_2b_1}{N}\right)^2 + \frac{N}{N_2N_B}\left(\frac{a_1b_2 - a_2b_1}{N}\right)^2$$

de donde, simplificando, se obtienen $$\chi^2 = \frac{N(a_1b_2 - a_2b_1)^2}{N_1N_2N_AN_B}$$

b) En el problema 12.14, $a_1 = 75$, $a_2 = 25$, $b_1 = 65$, $b_2 = 35$, $N_1 = 140$, $N_2 = 60$, $N_A = 100$, $N_B = 100$ y $N = 200$; entonces, como se obtuvo antes,

$$\chi^2 = \frac{200[(75)(35) - (25)(65)]^2}{(140)(60)(100)(100)} = 2.38$$

Empleando la corrección de Yates se llega al mismo resultado que en el problema 12.15:

$$\chi^2(\text{corregida}) = \frac{N(|a_1b_2 - a_2b_1| - \frac{1}{2}N)^2}{N_1N_2N_AN_B} = \frac{200[|(75)(35) - (25)(65)| - 100]^2}{(140)(60)(100)(100)} = 1.93$$

12.20 A 900 hombres y 900 mujeres se les preguntó si deseaban que hubiera más programas federales de ayuda para el cuidado de los niños. Cuarenta por ciento de las mujeres y 36 por ciento de los hombres respondieron que sí. Probar con $\alpha = 0.05$ la hipótesis nula de estos porcentajes iguales contra la hipótesis alternativa de estos porcentajes diferentes. Mostrar que la prueba ji cuadrada para dos proporciones muestrales es equivalente a la prueba de significancia para diferencias empleando la aproximación normal del capítulo 10.

SOLUCIÓN

Bajo la hipótesis H_0,

$$\mu_{P_1 - P_2} = 0 \text{ y } \sigma_{P_1 - P_2} = \sqrt{pq\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} = \sqrt{(0.38)(0.62)\left(\frac{1}{900} + \frac{1}{900}\right)} = 0.0229$$

donde p se estima fusionando las proporciones de las dos muestras. Es decir,

$$p = \frac{360 + 324}{900 + 900} = 0.38 \quad \text{y} \quad q = 1 - 0.38 = 0.62$$

El estadístico de prueba para la aproximación normal es el siguiente:

$$Z = \frac{P_1 - P_2}{\sigma_{P_1 - P_2}} = \frac{0.40 - 0.36}{0.0229} = 1.7467$$

El resultado que da MINITAB del análisis ji cuadrada es el siguiente:

Prueba ji cuadrada

```
Los resultados esperados se muestran debajo de los observados

        males     females    Total
1       324       360        684
        342.00    342.00
2       576       549        1 116
        558.00    558.00
Total   900       900        1 800

Ji-Sq = 0.947 + 0.947 +
        0.581 + 0.581 = 3.056
DF = 1, P-Value = 0.080
```

El cuadrado del estadístico de prueba normal es $(1.7467)^2 = 3.056$, que es el valor del estadístico ji cuadrada. Las dos pruebas son equivalentes. Los valores p son siempre los mismos para las dos pruebas.

COEFICIENTE DE CONTINGENCIA

12.21 Encontrar el coeficiente de contingencia correspondiente a los datos de la tabla de contingencia del problema 12.14.

SOLUCIÓN

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{2.38}{2.38 + 200}} = \sqrt{0.01176} = 0.1084$$

12.22 Encontrar el valor máximo de C correspondiente a la tabla 2×2 del problema 12.14.

SOLUCIÓN

El valor máximo de C se presenta cuando las dos clasificaciones son perfectamente dependientes o están muy bien relacionadas. En ese caso, todos los que toman el suero sanan y todos los que no lo toman no sanan. La tabla de contingencia, entonces, será como la tabla 12.18.

Tabla 12.18

	Sanados	No sanados	Total
Grupo A (usan el suero)	100	0	100
Grupo B (no usan el suero)	0	100	100
Total	100	100	200

Dado que las frecuencias de celda esperadas, suponiendo completa independencia, son todas igual a 50,

$$\chi^2 = \frac{(100-50)^2}{50} + \frac{(0-50)^2}{50} + \frac{(0-50)^2}{50} + \frac{(100-50)^2}{50} = 200$$

Por lo tanto, el máximo valor de C es $\sqrt{\chi^2/(\chi^2 + N)} = \sqrt{200/(200 + 200)} = 0.7071$.

En general, para que exista dependencia perfecta en una tabla de contingencia en la que la cantidad de renglones y de columnas son ambas igual a k, las únicas frecuencias de celda distintas de cero deben encontrarse en la diagonal que va de la esquina superior izquierda a la esquina inferior derecha de la tabla de contingencia. En tales casos, $C_{\text{máx}} = \sqrt{(k-1)/k}$. (Ver los problemas 12.52 y 12.53.)

CORRELACIÓN DE ATRIBUTOS

12.23 Encontrar el coeficiente de correlación correspondiente a la tabla 12.12 del problema 12.14: *a*) sin corrección de Yates y *b*) con corrección.

SOLUCIÓN

a) Como $\chi^2 = 2.28$, $N = 200$ y $k = 2$, se tiene

$$r = \sqrt{\frac{\chi^2}{N(k-1)}} = \sqrt{\frac{2.38}{200}} = 0.1091$$

lo que indica una correlación muy pequeña entre la recuperación de la salud y el uso del suero.

b) De acuerdo con el problema 12.15, r (corregida) $= \sqrt{1.93/200} = 0.0982$.

12.24 Demostrar que el coeficiente de correlación para tablas de contingencia, definido por la ecuación (*12*) de este capítulo, se encuentra entre 0 y 1.

SOLUCIÓN

De acuerdo con el problema 12.53, el valor máximo de $\sqrt{\chi^2/(\chi^2+N)}$ es $\sqrt{(k-1)/k}$. Por lo tanto,

$$\frac{\chi^2}{\chi^2+N} \le \frac{k-1}{k} \qquad k\chi^2 \le (k-1)(\chi^2+N) \quad k\chi^2 \le k\chi^2 - \chi^2 + kN - N$$

$$\chi^2 \le (k-1)N \qquad \frac{\chi^2}{N(k-1)} \le 1 \qquad \text{y} \qquad r = \sqrt{\frac{\chi^2}{N(k-1)}} \le 1$$

Como $\chi^2 \ge 0$, $r \ge 0$. Por lo tanto, $0 \le r \le 1$, que es lo requerido.

PROPIEDAD ADITIVA DE χ^2

12.25 Para probar una hipótesis H_0, se repite un experimento tres veces. Los valores que se obtienen para χ^2 son 2.37, 2.86 y 3.54, cada uno de los cuales corresponde a 1 grado de libertad. Mostrar que aunque no puede rechazarse H_0, al nivel 0.05, con base en ninguno de estos experimentos, sí puede rechazarse fusionando los tres experimentos.

SOLUCIÓN

El valor de χ^2 que se obtiene fusionando los resultados de los tres experimentos es, de acuerdo con la *propiedad aditiva*, $\chi^2 = 2.37 + 2.86 + 3.54 = 8.77$ con $1 + 1 + 1 = 3$ grados de libertad. Como $\chi^2_{.95}$ para 3 grados de libertad es 7.81, se puede rechazar H_0 al nivel de significancia 0.05. Pero como para 1 grado de libertad $\chi^2_{.95} = 3.84$, basándose en cualquiera de los tres experimentos, no se puede rechazar H_0.

Cuando se fusionan experimentos en los que se han obtenido valores de χ^2 que corresponden a 1 grado de libertad, se omite la corrección de Yates debido a que ésta tiende a sobrecorregir.

PROBLEMAS SUPLEMENTARIOS

LA PRUEBA JI CUADRADA

12.26 En 60 lanzamientos de una moneda se obtuvieron 37 caras y 23 cruces. Empleando como niveles de significancia: *a*) 0.05 y *b*) 0.01, probar la hipótesis de que la moneda no está cargada.

12.27 Resolver el problema 12.26 empleando la corrección de Yates.

12.28 Durante algún tiempo, las puntuaciones dadas a los alumnos por un grupo de profesores de determinada materia fueron, en promedio: 12% Aes; 18% Bes; 40% Ces; 18% Des, y 12% Efes. Durante dos semestres, un profesor nuevo da 22 Aes, 34 Bes, 66 Ces, 16 Des y 12 Efes. Al nivel de significancia 0.05, determinar si el nuevo profesor sigue el patrón de calificaciones establecido por los otros profesores.

12.29 Tres monedas se lanzan 240 veces anotando cada vez la cantidad de caras y de cruces que se obtienen. En la tabla 12.19 se muestran los resultados junto con los resultados esperados bajo la hipótesis de que las monedas no están cargadas. Probar esta hipótesis al nivel de significancia 0.05.

Tabla 12.19

	0 caras	1 cara	2 caras	3 caras
Frecuencias observadas	24	108	95	23
Frecuencias esperadas	30	90	90	30

12.30 En la tabla 12.20 se muestra el número de libros prestados en una biblioteca pública a lo largo de determinada semana. Probar la hipótesis de que el número de libros que se prestan no depende del día de la semana; usar los niveles de significancia: *a*) 0.05 y *b*) 0.01.

Tabla 12.20

	Lunes	Martes	Miércoles	Jueves	Viernes
Cantidad de libros prestados	135	108	120	114	146

12.31 Una urna contiene 6 canicas rojas y 3 canicas blancas. Se sacan en forma aleatoria dos canicas de la urna, se anotan sus colores y se devuelven a la urna. Este proceso se realiza 120 veces, los resultados obtenidos se presentan en la tabla 12.21.

a) Determinar las frecuencias esperadas.

b) Al nivel de significancia 0.05, determinar si los resultados obtenidos son consistentes con los resultados esperados.

Tabla 12.21

	0 rojas 2 blancas	1 roja 1 blanca	2 rojas 0 blancas
Número de extracciones	6	53	61

12.32 Se toman en forma aleatoria 200 pernos de la producción de cada una de cuatro máquinas. La cantidad de pernos defectuosos que se encuentran es 2, 9, 10 y 3. Empleando como nivel de significancia 0.05, determinar si hay una diferencia significativa entre las máquinas.

BONDAD DE AJUSTE

12.33 *a*) Emplear la prueba ji cuadrada para determinar la bondad de ajuste de los datos de la tabla 7.9 del problema 7.75. *b*) ¿Es el ajuste "demasiado bueno"? Emplear el nivel de significancia 0.05.

12.34 Usar la prueba ji cuadrada para determinar la bondad de ajuste de los datos: *a*) de la tabla 3.8 del problema 3.59 y *b*) de la tabla 3.10 del problema 3.61. Usar el nivel de significancia 0.05 y determinar, en cada caso, si el ajuste es "demasiado bueno".

12.35 Usar la prueba ji cuadrada para determinar la bondad de ajuste de los datos: *a*) de la tabla 7.9 del problema 7.75 y *b*) de la tabla 7.12 del problema 7.80 ¿Es el resultado obtenido en *a*) consistente con el del problema 12.33?

TABLAS DE CONTINGENCIA

12.36 La tabla 12.22 muestra el resultado de un experimento para investigar el efecto que tiene la vacunación contra determinada enfermedad en los animales de laboratorio. Empleando el nivel de significancia: *a*) 0.01 y *b*) 0.05, probar la hipótesis de que no hay diferencia entre el grupo vacunado y el no vacunado (es decir, la vacunación y la enfermedad son independientes).

Tabla 12.22

	Adquirieron la enfermedad	No adquirieron la enfermedad
Vacunados	9	42
No vacunados	17	28

Tabla 12.23

	Aprobaron	No aprobaron
Grupo *A*	72	17
Grupo *B*	64	23

12.37 Resolver el problema 12.36 empleando la corrección de Yates.

12.38 En la tabla 12.23 se presenta la cantidad de estudiantes de dos grupos, *A* y *B*, que aprobaron y que no aprobaron un examen realizado a ambos grupos. Empleando el nivel de significancia: *a*) 0.05 y *b*) 0.01, probar la hipótesis de que no hay diferencia entre los dos grupos. Resolver el problema con corrección de Yates y sin ella.

12.39 De un grupo de pacientes que se quejaba de no dormir bien, a algunos se les dieron unas pastillas para dormir, en tanto que a otros se les dieron pastillas de azúcar (aunque todos *pensaban* que se les daban pastillas para dormir). Después se les interrogó acerca de si las pastillas les habían ayudado a dormir o no. En la tabla 12.24 se muestran los resultados obtenidos. Suponiendo que todos los pacientes digan la verdad, probar la hipótesis de que no hay diferencia entre las pastillas para dormir y las pastillas de azúcar, empleando como nivel de significancia 0.05.

Tabla 12.24

	Durmió bien	No durmió bien
Tomó pastillas para dormir	44	10
Tomó pastillas de azúcar	81	35

12.40 En relación con determinada propuesta de interés nacional, los votos de demócratas y republicanos son como se muestra en la tabla 12.25. Al nivel de significancia: *a*) 0.01 y *b*) 0.05, probar la hipótesis de que, en lo referente a esta propuesta, no hay diferencia entre los dos partidos.

Tabla 12.25

	A favor	En contra	Indeciso
Demócratas	85	78	37
Republicanos	118	61	25

12.41 En la tabla 12.26 se muestra la relación que hay entre el desempeño de los estudiantes en matemáticas y en física. Probar la hipótesis de que el desempeño en matemáticas es independiente del desempeño en física, empleando el nivel de significancia: *a*) 0.05 y *b*) 0.01.

Tabla 12.26

		Matemáticas		
		Calificación alta	Calificación intermedia	Calificación baja
Física	Calificación alta	56	71	12
	Calificación intermedia	47	163	38
	Calificación baja	14	42	85

12.42 En la tabla 12.27 se muestran los resultados de una encuesta realizada con objeto de determinar si la edad de un conductor de 21 años o más tiene alguna relación con la cantidad de accidentes automovilísticos en los que se ve implicado (incluyendo accidentes menores). Al nivel de significancia: *a*) 0.05 y *b*) 0.01, probar la hipótesis de que la cantidad de accidentes es independiente de la edad del conductor. ¿Cuáles pueden ser las fuentes de dificultad en la técnica de muestreo, así como otras consideraciones, que puedan afectar los resultados?

Tabla 12.27

		Edad del conductor				
		21-30	31-40	41-50	51-60	61-70
Número de accidentes	0	748	821	786	720	672
	1	74	60	51	66	50
	2	31	25	22	16	15
	>2	9	10	6	5	7

12.43 *a*) Probar que $\chi^2 = \sum (\sigma_j^2/e_j) - N$ para todas las tablas de contingencia, donde N es la frecuencia total de todas las celdas.

b) Resolver el problema 12.41 empleando los resultados de *a*).

12.44 Si N_i y N_j denotan, respectivamente, la suma de las frecuencias en el renglón i y en la columna j de una tabla de contingencia (las *frecuencias marginales*), demostrar que la frecuencia esperada en la celda del renglón i y la columna j es $N_i N_j / N$, donde N es la frecuencia total de todas las celdas.

12.45 Probar la fórmula (*9*) de este capítulo. (*Sugerencia*: Utilizar los problemas 12.43 y 12.44.)

12.46 Extender la fórmula (*9*) de este capítulo a tablas de contingencia $2 \times k$, donde $k > 3$.

12.47 Probar la fórmula (*8*) de este capítulo.

12.48 Por analogía con las ideas desarrolladas para tablas de contingencia $h \times k$, analizar las tablas de contingencia $h \times k \times 1$, indicando sus posibles aplicaciones.

COEFICIENTE DE CONTINGENCIA

12.49 En la tabla 12.28 se muestra la relación entre color de pelo y color de ojos encontrada en una muestra de 200 estudiantes.

a) Encontrar el coeficiente de contingencia sin corrección de Yates y con ella.

b) Comparar el resultado de *a*) con el coeficiente máximo de contingencia.

Tabla 12.28

		Color de pelo	
		Rubio	No rubio
Color de ojos	Azules	49	25
	No azules	30	96

12.50 Encontrar el coeficiente de contingencia correspondiente a los datos: *a*) del problema 12.36 y *b*) del problema 12.38, con corrección de Yates y sin ella.

12.51 Encontrar el coeficiente de contingencia correspondiente a los datos del problema 12.41.

12.52 Probar que el coeficiente máximo de contingencia de una tabla 3×3 es $\sqrt{\frac{2}{3}} = 0.8165$, aproximadamente.

12.53 Probar que el coeficiente máximo de contingencia de una tabla $k \times k$ es $\sqrt{(k-1)/k}$.

CORRELACIÓN DE ATRIBUTOS

12.54 Encontrar el coeficiente de correlación de los datos de la tabla 12.28.

12.55 Encontrar el coeficiente de correlación de los datos: *a*) de la tabla 12.22 y *b*) de la tabla 12.23, con corrección de Yates y sin ella.

12.56 Encontrar el coeficiente de correlación entre las calificaciones de matemáticas y de física de la tabla 12.26.

12.57 Si C es el coeficiente de contingencia de una tabla $k \times k$ y r es el coeficiente de correlación correspondiente, probar que $r = C/\sqrt{(1 - C^2)(k-1)}$.

PROPIEDAD ADITIVA DE χ^2

12.58 Para probar una hipótesis H_0, se repite un experimento cinco veces. Los valores obtenidos para χ^2, correspondiente cada uno a 4 grados de libertad, son 8.3, 9.1, 8.9, 7.8 y 8.6. Mostrar que aunque al nivel de significancia 0.05 no se puede rechazar H_0 con base en ninguno de los experimentos por separado, sí se puede rechazar a este nivel de significancia con base en todos los experimentos juntos.

AJUSTE DE CURVAS Y MÉTODO DE MÍNIMOS CUADRADOS

13

RELACIÓN ENTRE VARIABLES

Con frecuencia, en la práctica se encuentra que existen relaciones entre dos (o más) variables. Por ejemplo, el peso de los hombres adultos depende de alguna manera de su estatura; la circunferencia de un círculo depende de su radio, y la presión de una masa de gas depende de su temperatura y volumen.

Es útil expresar estas relaciones en forma matemática mediante una ecuación que conecte estas variables.

AJUSTE DE CURVAS

Para hallar una ecuación que relacione las variables, el primer paso es obtener datos que muestren los valores de las variables que se están considerando. Por ejemplo, si X y Y denotan, respectivamente, la estatura y el peso de hombres adultos, entonces en una muestra de N individuos se hallan las estaturas $X_1, X_2, \ldots, X_N$ y los correspondientes pesos $Y_1, Y_2, \ldots, Y_N$.

El paso siguiente es graficar los puntos $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ en un sistema de coordenadas rectangulares. Al conjunto de puntos obtenido se le llama *diagrama de dispersión*.

En el diagrama de dispersión es posible visualizar alguna curva cuya forma se aproxime a los datos. A esta curva se le llama *curva de aproximación*. Por ejemplo, en la figura 13-1 los datos al parecer se aproximan adecuadamente mediante una línea recta; entonces se dice que entre las variables existe una relación lineal. En cambio, en la figura 13-2, aunque existe una relación entre las variables, esta relación no es una relación lineal y por lo tanto se le llama *relación no lineal*.

En general, al problema de hallar la ecuación de una curva de aproximación que se ajuste a un conjunto dado de datos se le conoce como *ajuste de curvas*.

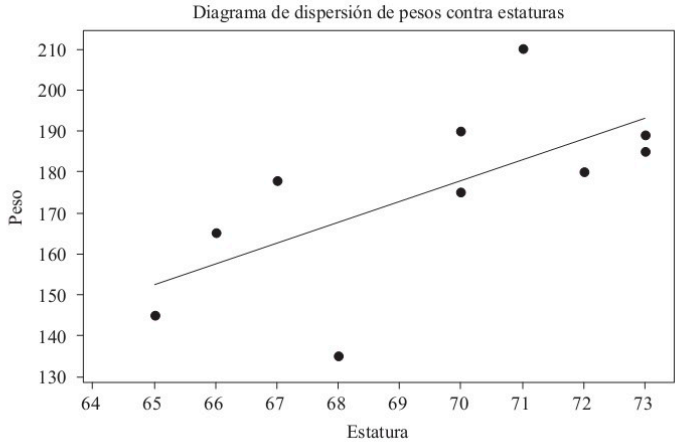


Figura 13-1 Algunas veces la relación entre dos variables se describe mediante una línea recta.

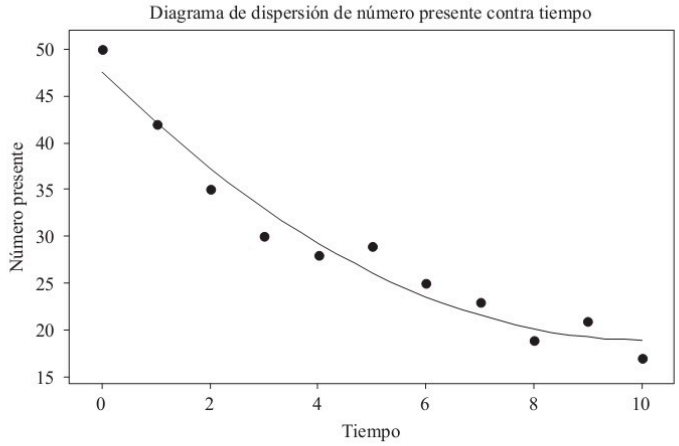


Figura 13-2 Algunas veces la relación entre dos variables se describe mediante una relación no lineal.

ECUACIONES DE CURVAS DE APROXIMACIÓN

Como referencia, a continuación se presentan varios de los tipos más comunes de curvas de aproximación. Todas las letras, excepto X y Y, representan constantes. A las variables X y Y se les llama *variable independiente* y *variable dependiente*, respectivamente, aunque estos papeles pueden intercambiarse.

Línea recta	$Y = a_0 + a_1 X$	(*1*)
Parábola o curva cuadrática	$Y = a_0 + a_1 X + a_2 X^2$	(*2*)
Curva cúbica	$Y = a_0 + a_1 X + a_2 X^2 + a_3 X^3$	(*3*)
Curva cuártica	$Y = a_0 + a_1 X + a_2 X^2 + a_3 X^3 + a_4 X^4$	(*4*)
Curva de grado n	$Y = a_0 + a_1 X + a_2 X^2 + \cdots + a_n X^n$	(*5*)

En las ecuaciones anteriores, a las expresiones de los lados derechos se les conoce como *polinomios* de primero, segundo, tercero, cuarto y *n*-ésimo grados, respectivamente. Las funciones definidas por las primeras cuatro ecuaciones se llaman funciones *lineales*, *cuadráticas*, *cúbicas* y *cuárticas*, en ese orden.

Las siguientes son algunas de las muchas otras funciones que se emplean en la práctica:

Hipérbola	$Y = \dfrac{1}{a_0 + a_1 X}$ o bien $\dfrac{1}{Y} = a_0 + a_1 X$	(6)
Curva exponencial	$Y = ab^X$ o bien $\log Y = \log a + (\log b)X = a_0 + a_1 X$	(7)
Curva geométrica	$Y = aX^b$ o bien $\log Y = \log a + b(\log X)$	(8)
Curva exponencial modificada	$Y = ab^X + g$	(9)
Curva geométrica modificada	$Y = aX^b + g$	(10)
Curva de Gompertz	$Y = pq^{b^X}$ o bien $\log Y = \log p + b^X(\log q) = ab^X + g$	(11)
Curva de Gompertz modificada	$Y = pq^{b^X} + h$	(12)
Curva logística	$Y = \dfrac{1}{ab^X + g}$ o bien $\dfrac{1}{Y} = ab^X + g$	(13)
	$Y = a_0 + a_1(\log X) + a_2(\log X)^2$	(14)

Para saber cuál de estas curvas emplear, es útil obtener el diagrama de dispersión de las variables transformadas. Por ejemplo, si el diagrama de dispersión de log Y contra X muestra una relación lineal, la ecuación será de la forma (7), en tanto que si log Y contra log X muestra una relación lineal, la ecuación será de la forma (8). Como ayuda para saber qué tipo de curva utilizar suele emplearse papel especial para graficar. Al papel para graficar en el que una de las escalas está calibrada logarítmicamente se le conoce como *papel semilogarítmico*, y al papel en el que las dos escalas están calibradas de manera logarítmica se le conoce como *papel logarítmico*.

MÉTODO DE AJUSTE DE CURVAS A MANO

Para trazar una curva de aproximación que se ajuste a los datos puede emplearse el criterio personal. A este método se le llama *ajuste de curva a mano*. Si se sabe cuál es el tipo de ecuación, las constantes de la ecuación se determinan eligiendo tantos puntos de la curva como constantes tenga la ecuación. Por ejemplo, si la curva es una línea recta, se necesitarán dos puntos; si es una parábola, se necesitarán tres puntos. Este método tiene la desventaja de que personas distintas encontrarán curvas y ecuaciones distintas.

LA LÍNEA RECTA

El tipo más sencillo de curva de aproximación es una línea recta, cuya ecuación puede escribirse como

$$Y = a_0 + a_1 X \tag{15}$$

Dados dos puntos cualesquiera (X_1, Y_1) y (X_2, Y_2) de la recta, se determinan las constantes a_0 y a_1. La ecuación que se obtiene es

$$Y - Y_1 = \left(\frac{Y_2 - Y_1}{X_2 - X_1}\right)(X - X_1) \qquad \text{o bien} \qquad Y - Y_1 = m(X - X_1) \tag{16}$$

donde

$$m = \frac{Y_2 - Y_1}{X_2 - X_1}$$

es la *pendiente* de la recta y representa el cambio o variación en Y dividido por un cambio o variación correspondiente en X.

En la ecuación escrita de la forma (*15*), la constante a_1 es la pendiente m. La constante a_0, que es el valor de Y cuando $X = 0$, se conoce como la *intersección con el eje Y*.

EL MÉTODO DE MÍNIMOS CUADRADOS

Para evitar el empleo del criterio personal para la construcción de rectas, parábolas u otras curvas de aproximación que se ajusten a un conjunto de datos, es necesario ponerse de acuerdo en una definición de la "recta de mejor ajuste", la "parábola de mejor ajuste", etcétera.

Con objeto de dar una definición, considérese la figura 13-3, en la que los datos son los puntos $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$. Dado un valor de X, por ejemplo X_1, entre el valor Y_1 y el valor correspondiente determinado de acuerdo con la curva C habrá una diferencia.Como se muestra en la figura, esta diferencia se denota D_1 y se llama la *desviación*, el *error* o el *residual* y puede ser positivo, negativo o cero. De manera semejante se obtienen las desviaciones $X_2, \ldots, X_N$ correspondientes a cada valor $D_2, \ldots, D_N$.

Una medida de la "bondad de ajuste" de la curva C a los datos dados es la cantidad $D_1^2 + D_2^2 + \cdots + D_N^2$. Si esta cantidad es pequeña, el ajuste es bueno; si es grande, el ajuste es malo. De esta manera se llega a la definición siguiente:

> **Definición:** De todas las curvas que se aproximan a un conjunto dado de puntos, a la curva que tiene la propiedad de que $D_1^2 + D_2^2 + \cdots + D_N^2$ sea la mínima se le llama *curva de mejor ajuste*.

Una curva que tiene esta propiedad se dice que se ajusta a los datos en el *sentido de mínimos cuadrados* y se le llama *curva de mínimos cuadrados*. De manera que una recta que tiene esta propiedad se dice que es una *recta de mínimos cuadrados*, una parábola que tiene esta propiedad es una *parábola de mínimos cuadrados*, etcétera.

La definición anterior suele emplearse cuando X es la variable independiente y Y es la variable dependiente. Si X es la variable dependiente, la definición se modifica considerando desviaciones horizontales en lugar de desviaciones verticales, lo que equivale a intercambiar los ejes X y Y. Por lo general, estas dos definiciones llevan a curvas distintas de mínimos cuadrados. En este libro, a menos que se especifique otra cosa, se considerará que X es la variable independiente y que Y es la variable dependiente.

También pueden definirse otras curvas de mínimos cuadrados considerando las distancias perpendiculares del punto a la curva en lugar de las distancias verticales u horizontales. Sin embargo, esto no suele usarse.

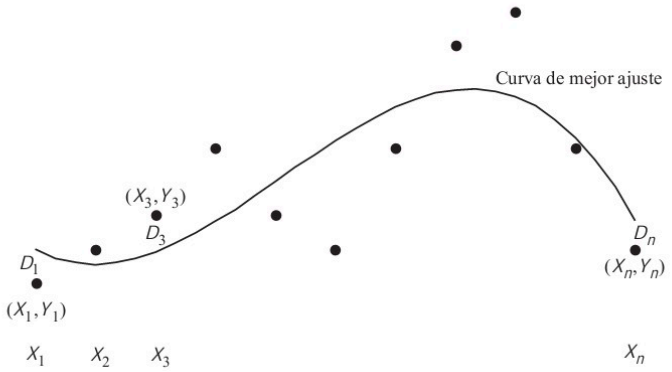


Figura 13-3 D_1 **es la distancia del punto** (X_1, Y_1) **a la curva de mejor ajuste,** $\ldots$, D_n **es la distancia del punto** (X_n, Y_n) **a la curva de mejor ajuste.**

LA RECTA DE MÍNIMOS CUADRADOS

La recta de mínimos cuadrados que aproxima el conjunto de puntos $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ tiene la ecuación

$$Y = a_0 + a_1 X \tag{17}$$

donde las constantes a_0 y a_1 se determinan resolviendo las ecuaciones simultáneas

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 \end{aligned} \tag{18}$$

a las que se les denomina *ecuaciones normales de la recta de mínimos cuadrados* (*17*). Las constantes a_0 y a_1 de las ecuaciones (*18*) pueden hallarse empleando las fórmulas

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2} \qquad a_1 = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} \tag{19}$$

Para recordar las ecuaciones normales (*18*) hay que observar que la primera ecuación se obtiene formalmente sumando a ambos lados de la ecuación (*17*) [es decir, $\sum Y = \sum (a_0 + a_1 X) = a_0 N + a_1 \sum X$] y la segunda ecuación se obtiene multiplicando, primero, ambos lados de la ecuación (*17*) por X y después sumando [es decir, $\sum XY = \sum X(a_0 + a_1 X) = a_0 \sum X + a_1 \sum X^2$]. Obsérvese que no se trata de una deducción de las ecuaciones normales, sino simplemente de una manera que facilita recordarlas. Obsérvese también que en las ecuaciones (*18*) y (*19*) se ha empleado la notación abreviada $\sum X$, $\sum XY$, etc., en lugar de $\sum_{j=1}^{N} X_j$, $\sum_{j=1}^{N} X_j Y_j$, etcétera.

El trabajo que implica hallar la recta de mínimos cuadrados puede reducirse transformando los datos de manera que $x = X - \bar{X}$ y $y = Y - \bar{Y}$. Entonces, la ecuación de la recta de mínimos cuadrados puede escribirse de la manera siguiente (problema 13.15):

$$y = \left(\frac{\sum xy}{\sum x^2}\right) x \qquad \text{o bien} \qquad y = \left(\frac{\sum xY}{\sum x^2}\right) x \tag{20}$$

En particular, si X es tal que $\sum X = 0$ (es decir, $\bar{X} = 0$), la ecuación se convierte en

$$Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right) X \tag{21}$$

La ecuación (*20*) implica que $y = 0$ para $x = 0$; por lo tanto, la recta de mínimos cuadrados pasa por el punto $(\bar{X}, \bar{Y})$, al que se le llama el *centroide* o *centro de gravedad* de los datos.

Si se considera que la variable X es la variable dependiente en lugar de la variable independiente, la ecuación (*17*) se escribe $X = b_0 + b_1 Y$. Las fórmulas anteriores también son válidas cuando se intercambian X y Y, y a_0 y a_1 se sustituyen por b_0 y b_1, respectivamente. Sin embargo, por lo general la recta de mínimos cuadrados que se obtiene no es la misma que la que se obtuvo antes [ver problemas 13.11 y 13.15*d*)].

RELACIONES NO LINEALES

Algunas veces, las relaciones no lineales pueden reducirse a relaciones lineales mediante transformaciones adecuadas de las variables (ver problema 13.21).

LA PARÁBOLA DE MÍNIMOS CUADRADOS

La parábola de mínimos cuadrados que aproxima el conjunto de puntos $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ tiene la ecuación

$$Y = a_0 + a_1 X + a_2 X^2 \tag{22}$$

donde las constantes a_0, a_1 y a_2 se determinan resolviendo simultáneamente las ecuaciones

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X + a_2 \sum X^2 \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum X^3 \\ \sum X^2 Y &= a_0 \sum X^2 + a_1 \sum X^3 + a_2 \sum X^4 \end{aligned} \tag{23}$$

llamadas *ecuaciones normales de la parábola de mínimos cuadrados* (*22*).

Para recordar las ecuaciones (*23*), obsérvese que se pueden obtener formalmente multiplicando la ecuación (*22*) por 1, X y X^2, respectivamente, y sumando a ambos lados de las ecuaciones resultantes. Esta técnica puede extenderse a las ecuaciones normales de curvas cúbicas de mínimos cuadrados, ecuaciones cuárticas de mínimos cuadrados y, en general, a cualquiera de las curvas de mínimos cuadrados correspondientes a la ecuación (*5*).

Como en el caso de la recta de mínimos cuadrados, las ecuaciones (*23*) se simplifican si las X se escogen de manera que $\sum X = 0$. Estas ecuaciones también se simplifican empleando las nuevas variables $x = X - \bar{X}$ y $y = Y - \bar{Y}$.

REGRESIÓN

Con frecuencia se desea estimar el valor de la variable Y que corresponde a un valor dado de la variable X, basándose en los datos muestrales. Esto se hace estimando el valor de Y a partir de la curva de mínimos cuadrados ajustada a los datos muestrales. A la curva de mínimos cuadrados se le llama *curva de regresión de Y en X*, debido a que Y se estima a partir de X.

Si lo que se desea es estimar un valor de X a partir de un valor dado de Y, se emplea la *curva de regresión de X en Y*, que es lo mismo que intercambiar las variables en el diagrama de dispersión, de manera que X sea la variable dependiente y Y sea la variable independiente. En este caso se sustituyen las desviaciones verticales, de la definición de la curva de mínimos cuadrados de la página 284, por desviaciones horizontales.

En general, la recta o la curva de regresión de Y en X no es igual a la recta o a la curva de regresión de X en Y.

APLICACIONES A SERIES DE TIEMPO

Si la variable independiente X representa tiempo, los datos dan el valor de Y en distintos momentos. A los datos ordenados de acuerdo con el tiempo se les llama *serie de tiempo*. En este caso, a la recta o a la curva de regresión de Y en X se le llama *recta* o *curva de tendencia* y se emplea para hacer *estimaciones*, *predicciones* o *pronósticos*.

PROBLEMAS EN LOS QUE INTERVIENEN MÁS DE DOS VARIABLES

Los problemas en los que intervienen más de dos variables se tratan de manera análoga a los problemas de dos variables. Por ejemplo, entre las tres variables X, Y y Z puede haber una relación que pueda ser descrita mediante la ecuación

$$Z = a_0 + a_1 X + a_2 Y \tag{24}$$

a la que se le llama *ecuación lineal en las variables X, Y y Z*.

En un sistema de coordenadas rectangulares, esta ecuación representa un plano y los puntos muestrales (X_1, Y_1, Z_1), $(X_2, Y_2, Z_2), \ldots, (X_N, Y_N, Z_N)$ estarán "dispersos" no demasiado lejos de este plano, al que se le llama *plano de aproximación*.

Por extensión del método de mínimos cuadrados, se puede hablar de un *plano de mínimos cuadrados* que se aproxime a los datos. Si Z se aproxima a partir de los valores de X y Y, a este plano se le llamará *plano de regresión de Z en X y Y*. Las ecuaciones normales correspondientes al plano de mínimos cuadrados (*24*) son

$$\begin{aligned} \sum Z &= a_0 N + a_1 \sum X + a_2 \sum Y \\ \sum XZ &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum XY \\ \sum YZ &= a_0 \sum Y + a_1 \sum XY + a_2 \sum Y^2 \end{aligned} \tag{25}$$

y para recordarlas se puede pensar que se obtienen a partir de la ecuación (*24*) multiplicando ésta por 1, X y Y y sumando después.

También se pueden considerar ecuaciones más complicadas que la (*24*). Éstas representan *superficies de regresión*. Cuando el número de variables es mayor a tres, se pierde la intuición geométrica debido a que se requieren espacios de cuatro, cinco o n dimensiones.

A los problemas en los que se estima una variable a partir de dos o más variables se les llama problemas de *regresión múltiple*. Estos problemas serán considerados más detalladamente en el capítulo 15.

PROBLEMAS RESUELTOS

LÍNEAS RECTAS

13.1 Treinta estudiantes de secundaria fueron entrevistados en un estudio acerca de la relación entre el tiempo que pasan en Internet y su promedio de calificaciones. Los resultados se muestran en la tabla 13.1. X es la cantidad de tiempo que pasan en Internet y Y es su promedio de calificaciones.

Tabla 13.1

Horas	Promedio	Horas	Promedio	Horas	Promedio
11	2.84	9	2.85	25	1.85
5	3.20	5	3.35	6	3.14
22	2.18	14	2.60	9	2.96
23	2.12	18	2.35	20	2.30
20	2.55	6	3.14	14	2.66
20	2.24	9	3.05	19	2.36
10	2.90	24	2.06	21	2.24
19	2.36	25	2.00	7	3.08
15	2.60	12	2.78	11	2.84
18	2.42	6	2.90	20	2.45

Usar MINITAB para:

a) Hacer un diagrama de dispersión con estos datos.

b) Ajustar una recta a estos datos y dar los valores de a_0 y a_1.

SOLUCIÓN

a) En las columnas `C1` y `C2` de la hoja de cálculo de MINITAB se ingresan estos datos. La columna `C1` se titula `Horas en Internet` y la columna `C2` `Promedio de calificaciones`. Empleando la secuencia **Stat** → **Regresión** → **Regression** se obtienen los resultados que se muestran en la figura 13-4.

b) El valor de a_0 es 3.49 y el valor de a_1 es −0.0594.

13.2 Resolver el problema 13.1 usando EXCEL.

SOLUCIÓN

En las columnas A y B de la hoja de cálculo de EXCEL se ingresan los datos. Con la secuencia **Tools** → **Data Análisis** → **Regression** se obtiene el cuadro de diálogo de la figura 13-5 que se llena como ahí se muestra. La parte de interés del resultado, en este momento, es

Intersección	3.488753
Horas en Internet	−0.05935

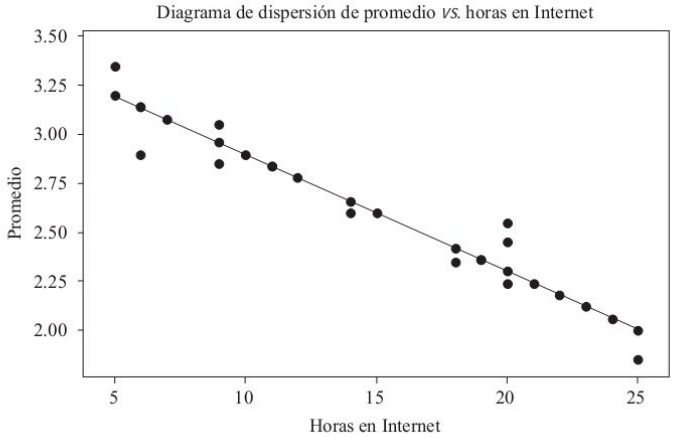


Figura 13-4 La suma de los cuadrados de las distancias de los puntos a la recta de mejor ajuste es la mínima utilizando la recta promedio = 3.49 – 0.0594 horas en Internet.

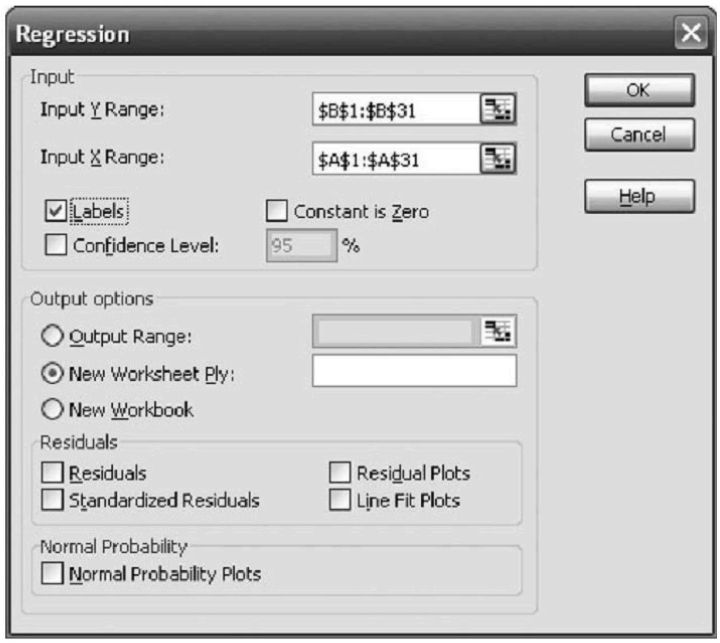


Figura 13-5 EXCEL, cuadro de diálogo para el problema 13.2.

A la constante a_0 se le llama *intersección* y a la constante a_1 se le denomina *pendiente*. Se obtienen los mismos valores que con MINITAB.

13.3 *a*) Mostrar que la ecuación de la recta que pasa a través de los puntos (X_1, Y_1) y (X_2, Y_2) está dada por

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1)$$

b) Encontrar la ecuación de la recta que pasa a través de los puntos (2, −3) y (4, 5).

SOLUCIÓN

a) La ecuación de la recta es

$$Y = a_0 + a_1 X \tag{29}$$

Como (X_1, Y_1) está en la recta,

$$Y_1 = a_0 + a_1 X_1 \tag{30}$$

Como (X_2, Y_2) está en la recta,

$$Y_2 = a_0 + a_1 X_2 \tag{31}$$

Sustrayendo la ecuación (*30*) de la ecuación (*29*),

$$Y - Y_1 = a_1(X - X_1) \tag{32}$$

Sustrayendo la ecuación (*30*) de la ecuación (*31*),

$$Y_2 - Y_1 = a_1(X_2 - X_1) \qquad \text{o bien} \qquad a_1 = \frac{Y_2 - Y_1}{X_2 - X_1}$$

Sustituyendo este valor de a_1 en la ecuación (*32*), se obtiene

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1)$$

como se deseaba. La cantidad

$$\frac{Y_2 - Y_1}{X_2 - X_1}$$

se abrevia *m*, representa el cambio en *Y* dividido entre el correspondiente cambio en *X* y es la *pendiente* de la recta. La ecuación buscada es $Y - Y_1 = m(X - X_1)$.

b) **Primer método** [empleando los resultados del inciso *a*)]

En el primer punto (2, −3) se tiene $X_1 = 2$ y $Y_1 = -3$; en el segundo punto (4, 5) se tiene $X_2 = 4$ y $Y_2 = 5$. Por lo tanto, la pendiente es

$$m = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{5 - (-3)}{4 - 2} = \frac{8}{2} = 4$$

y la ecuación buscada es

$$Y - Y_1 = m(X - X_1) \qquad \text{o bien} \qquad Y - (-3) = 4(X - 2)$$

la cual se puede escribir como $Y + 3 = 4(X - 2)$, o bien $Y = 4X - 11$.

Segundo método

La ecuación de una línea recta es $Y = a_0 + a_1 X$. Como el punto (2, −3) pertenece a esta recta, $-3 = a_0 + 2a_1$, y como también el punto (4, 5) pertenece a esta recta, $5 = a_0 + 4a_1$; resolviendo estas dos ecuaciones simultáneas, se obtiene $a_1 = 4$ y $a_0 = -11$. Por lo tanto, la ecuación buscada es

$$Y = -11 + 4X \qquad \text{o bien} \qquad Y = 4X - 11$$

13.4 Se siembra trigo en 9 parcelas del mismo tamaño. En la tabla 13.2 se muestran las cantidades de fertilizante empleadas en cada parcela, así como las cantidades de trigo obtenidas.

Usar MINITAB para ajustar una curva parabólica $Y = a_0 + a_1 X + a_2 X^2$ a estos datos.

Tabla 13.2

Cantidad de trigo (y)	Fertilizante (x)
2.4	1.2
3.4	2.3
4.4	3.3
5.1	4.1
5.5	4.8
5.2	5.0
4.9	5.5
4.4	6.1
3.9	6.9

SOLUCIÓN

Las cantidades de trigo se ingresan en la columna `C1` y las de fertilizante en la columna `C2`. Con la secuencia **Stat** → **Regresión** → **Fitted Line Plot** se obtiene el cuadro de diálogo que se muestra en la figura 13-6.

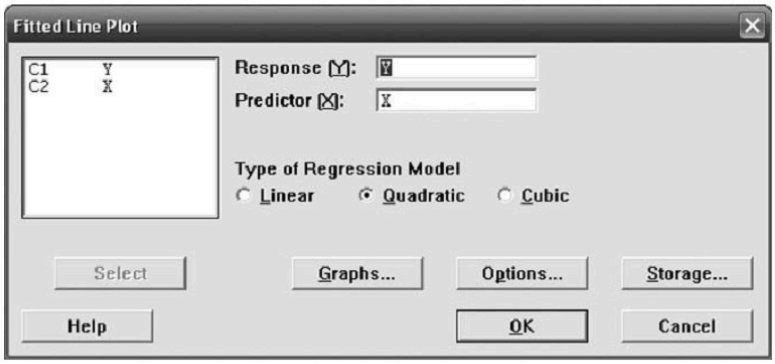


Figura 13-6 MINITAB, cuadro de diálogo para el problema 13.4.

Con este cuadro de diálogo se obtiene el resultado que se muestra en la figura 13-7.

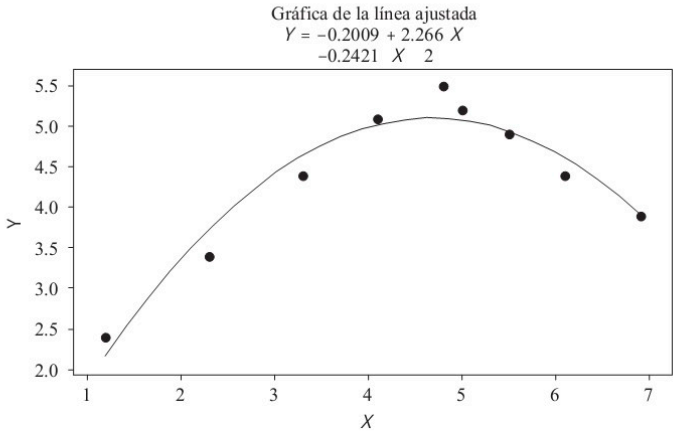


Figura 13-7 MINITAB, ajuste de la curva parabólica de mínimos cuadrados a un conjunto de datos.

13.5 Encontrar: *a*) la pendiente, *b*) la ecuación, *c*) la intersección con el eje Y y *d*) la intersección con el eje X de la recta que pasa por los puntos (1, 5) y (4, −1).

SOLUCIÓN

a) $(X_1 = 1, Y_1 = 5)$ y $(X_2 = 4, Y_2 = -1)$. Por lo tanto,

$$m = \text{pendiente} = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{-1 - 5}{4 - 1} = \frac{-6}{3} = -2$$

El signo negativo de la pendiente indica que a medida que X crece, Y decrece, como se muestra en la figura 13-8.

b) La ecuación de la recta es

$$Y - Y_1 = m(X - X_1) \qquad \text{o} \qquad Y - 5 = -2(X - 1)$$

Es decir, $$Y - 5 = -2X + 2 \qquad \text{o} \qquad Y = 7 - 2X$$

Esta ecuación también se puede obtener empleando el segundo método del problema 13.3*b*).

c) La intersección con el eje Y, que es el valor de Y cuando $X = 0$, es $Y = 7 - 2(0) = 7$. Esto también puede verse directamente en la figura 13-8.

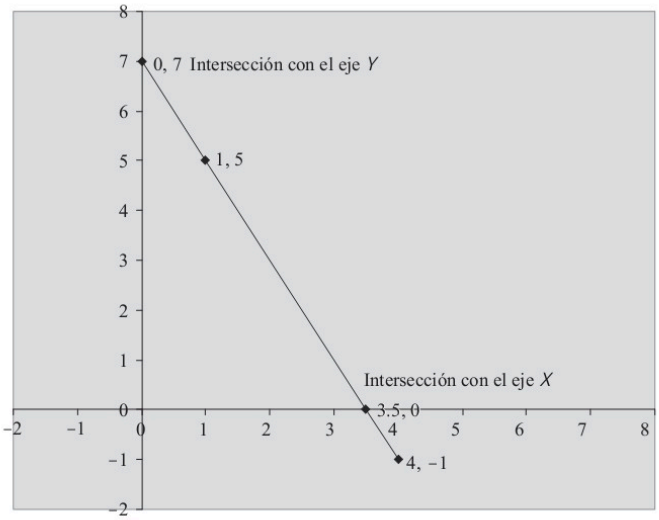


Figura 13-8 Recta que muestra la intersección con el eje X y la intersección con el eje Y.

d) La intersección con el eje X es el valor de X cuando $Y = 0$. Sustituyendo $Y = 0$ en la ecuación $Y = 7 - 2X$, se tiene $0 = 7 - 2X$, o $2X = 7$ y $X = 3.5$. Esto también se puede ver directamente en la figura 13-8.

13.6 Encontrar la ecuación de la recta que pasa a través del punto (4, 2) y que es paralela a la recta $2X + 3Y = 6$.

SOLUCIÓN

Si dos rectas son paralelas, sus pendientes son iguales. De $2X + 3Y = 6$ se obtiene $3Y = 6 - 2X$, o bien $Y = 2 - \frac{2}{3}X$, de manera que la pendiente de la recta es $m = -\frac{2}{3}$. Por lo tanto, la ecuación de la recta que se busca es

$$Y - Y_1 = m(X - X_1) \qquad \text{o} \qquad Y - 2 = -\tfrac{2}{3}(X - 4)$$

la cual también se puede escribir como $2X + 3Y = 14$.

Otro método

La ecuación de cualquier recta paralela a $2X + 3Y = 6$ es de la forma $2X + 3Y = c$. Para encontrar c, sea $X = 4$ y $Y = 2$. Entonces $2(4) + 3(2) = c$, o $c = 14$, con lo que la ecuación buscada es $2X + 3Y = 14$.

13.7 Encontrar la ecuación de la recta cuya pendiente es -4 y cuya intersección con el eje Y es 16.

SOLUCIÓN

En la ecuación $Y = a_0 + a_1X$, $a_0 = 16$ es la intersección con el eje Y y $a_1 = -4$ es la pendiente. Por lo tanto, la ecuación buscada es $Y = 16 - 4X$.

13.8 *a*) Construir una recta que se aproxime a los datos de la tabla 13.3.
b) Encontrar la ecuación de esta recta.

Tabla 13.3

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

SOLUCIÓN

a) En un sistema de coordenadas rectangulares se grafican los puntos (1, 1), (3, 2), (4, 4), (6, 4), (8, 5), (9, 7), (11, 8) y (14, 9), como se muestra en la figura 13-9. En la figura se ha trazado *a mano* una recta que se aproxima a los datos. En el problema 13.11 se muestra un método que elimina el criterio personal; ese método es el de mínimos cuadrados.

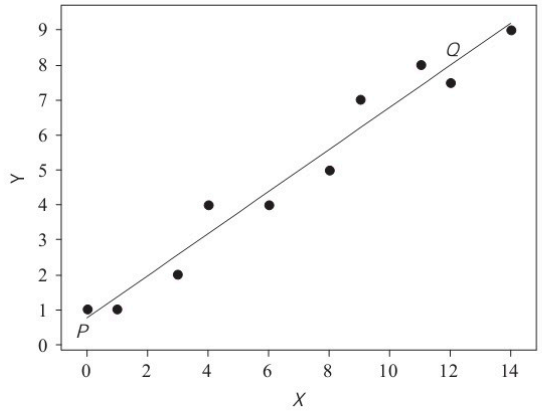


Figura 13-9 Método a mano para el ajuste de curvas.

b) Para obtener la ecuación de la recta construida en el inciso *a*), se eligen cualesquiera dos puntos de la recta, por ejemplo, P y Q; como se muestra en la gráfica, las coordenadas de los puntos P y Q son aproximadamente (0, 1) y (12, 7.5). La ecuación de una recta es $Y = a_0 + a_1X$. Por lo tanto, para el punto (0, 1) se tiene $1 = a_0 + a_1(0)$, y para el punto (12, 7.5) se tiene $7.5 = a_0 + 12a_1$; como de la primera de estas ecuaciones se obtiene $a_0 = 1$, de la segunda se obtiene $a_1 = 6.5/12 = 0.542$. Entonces, la ecuación buscada es $Y = 1 + 0.542X$.

Otro método

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1) \qquad \text{y} \qquad Y - 1 = \frac{7.4 - 1}{12 - 0}(X - 0) = 0.542X$$

Por lo tanto, $Y = 1 + 0.542X$.

13.9 *a*) Comparar los valores de Y obtenidos a partir de la recta de aproximación con los datos de la tabla 13.2.
b) Estimar el valor de Y para $X = 10$.

SOLUCIÓN

a) Para $X = 1$, $Y = 1 + 0.542(1) = 1.542$, o bien 1.5. Para $X = 3$, $Y = 1 + 0.542(3) = 2.626$ o bien 2.6. De la misma manera se obtienen valores de Y correspondientes a otros valores de X. Los valores estimados para Y a partir de la ecuación $Y = 1 + 0.542X$ se denotan Y_{est}. En la tabla 13.4 se presentan estos valores estimados junto con los datos originales.

b) El valor estimado de Y correspondiente a $X = 10$ es $Y = 1 + 0.542(10) = 6.42$ o 6.4.

Tabla 13.4

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9
Y_{est}	1.5	2.6	3.2	4.3	5.3	5.9	7.0	8.6

13.10 En la tabla 13.5 se presentan las estaturas en pulgadas (in) y los pesos en libras (lb) de 12 estudiantes varones que forman una muestra aleatoria de los estudiantes de primer año de una universidad.

Tabla 13.5

Estatura X (in)	70	63	72	60	66	70	74	65	62	67	65	68
Peso Y (lb)	155	150	180	135	156	168	178	160	132	145	139	152

a) Obtener el diagrama de dispersión de estos datos.
b) Trazar una recta que se aproxime a los datos.
c) Encontrar la ecuación de la recta que se trazó en el inciso *b*).
d) Estimar el peso de un estudiante cuya estatura es 63 in.
e) Estimar la estatura de un estudiante cuyo peso es 168 lb.

SOLUCIÓN

a) El diagrama de dispersión que se muestra en la figura 13-10 se obtiene graficando los puntos (70, 155), (63, 150),..., (68, 152).

b) En la figura 13-10 se presenta una recta que se aproxima a los datos. Pero ésta es sólo una de las muchas que podían haberse trazado.

c) Se toman dos puntos cualesquiera de la recta construida en el inciso *b*), por ejemplo P y Q. Las coordenadas de estos puntos, de acuerdo con la gráfica, son aproximadamente (60, 130) y (72, 170). Por lo tanto,

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1) \qquad Y - 130 = \frac{170 - 130}{72 - 60}(X - 60) \qquad Y = \frac{10}{3}X - 70$$

d) Si $X = 63$, entonces $Y = \frac{10}{3}(63) - 70 = 140$ lb.

e) Si $Y = 168$, entonces $168 = \frac{10}{3}X - 70$, $\frac{10}{3}X = 238$ y $X = 71.4$ o bien 71 in.

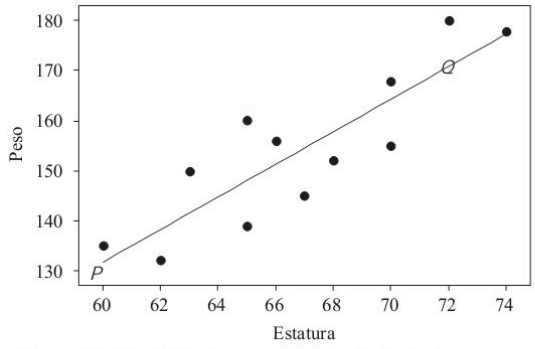


Figura 13-10 Método a mano para el ajuste de curvas.

LA RECTA DE MÍNIMOS CUADRADOS

13.11 Encontrar la recta de mínimos cuadrados correspondiente a los datos del problema 13.8 empleando: *a*) X como variable independiente y *b*) X como variable dependiente.

SOLUCIÓN

a) La ecuación de una recta es $Y = a_0 + a_1X$. Las ecuaciones normales son

$$\sum Y = a_0N + a_1 \sum X$$
$$\sum XY = a_0 \sum X + a_1 \sum X^2$$

El cálculo de estas sumas se puede organizar como se muestra en la tabla 13.6. Aunque la última columna de la derecha no se necesita en esta parte del problema, se ha incluido en la tabla para emplearla en el inciso *b*).

Como hay ocho pares de valores X y Y, $N = 8$ y las ecuaciones normales resultan ser

$$8a_0 + 56a_1 = 40$$
$$56a_0 + 524a_1 = 364$$

Resolviendo simultáneamente estas ecuaciones, se obtiene $a_0 = \frac{6}{11}$ o 0.545; $a_1 = \frac{7}{11}$ o 0.636; con lo que la recta de mínimos cuadrados buscada es $Y = \frac{6}{11} + \frac{7}{11}X$, o $Y = 0.545 + 0.636X$.

Tabla 13.6

X	Y	X^2	XY	Y^2
1	1	1	1	1
3	2	9	6	4
4	4	16	16	16
6	4	36	24	16
8	5	64	40	25
9	7	81	63	49
11	8	121	88	64
14	9	196	126	81
$\sum X = 56$	$\sum Y = 40$	$\sum X^2 = 524$	$\sum XY = 364$	$\sum Y^2 = 256$

Otro método

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N\sum X^2 - (\sum X)^2} = \frac{(40)(524) - (56)(364)}{(8)(524) - (56)^2} = \frac{6}{11} \quad \text{o bien} \quad 0.545$$

$$a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = \frac{(8)(364) - (56)(40)}{(8)(524) - (56)^2} = \frac{7}{11} \quad \text{o bien} \quad 0.636$$

Por lo tanto, $Y = a_0 + a_1X$, o bien $Y = 0.545 + 0.636X$, como antes.

b) Si X es considerada como la variable dependiente, entonces Y es la variable independiente; la ecuación de la recta de mínimos cuadrados es $X = b_0 + b_1Y$ y las ecuaciones normales son

$$\sum X = b_0 N + b_1 \sum Y$$

$$\sum XY = b_0 \sum Y + b_1 \sum Y^2$$

Entonces, de acuerdo con la tabla 13.6, las ecuaciones normales son

$$8b_0 + 40b_1 = 56$$

$$40b_0 + 256b_1 = 364$$

de donde $b_0 = -\frac{1}{2}$ o bien -0.50 y $b_1 = \frac{3}{2}$ o bien 1.50. Estos valores también pueden obtenerse de la manera siguiente

$$b_0 = \frac{(\sum X)(\sum Y^2) - (\sum Y)(\sum XY)}{N\sum Y^2 - (\sum Y)^2} = \frac{(56)(256) - (40)(364)}{(8)(256) - (40)^2} = -0.50$$

$$b_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum Y^2 - (\sum Y)^2} = \frac{(8)(364) - (56)(40)}{(8)(256) - (40)^2} = 1.50$$

Por lo tanto, la ecuación buscada de la recta de mínimos cuadrados es $X = b_0 + b_1Y$ o bien $X = -0.50 + 1.50Y$.

Obsérvese que despejando Y de esta ecuación se obtiene $Y = \frac{1}{3} + \frac{2}{3}X$ o bien $Y = 0.333 + 0.667X$, que no es igual a la recta obtenida en el inciso *a*).

13.12 Emplear el paquete para estadística SAS para trazar, en una misma gráfica, los puntos correspondientes a los datos de estatura y peso del problema 13.10 y la recta de mínimos cuadrados.

SOLUCIÓN

En la figura 13-11, los puntos correspondientes a los datos se presentan como pequeños círculos vacíos y la recta de mínimos cuadrados como una recta punteada.

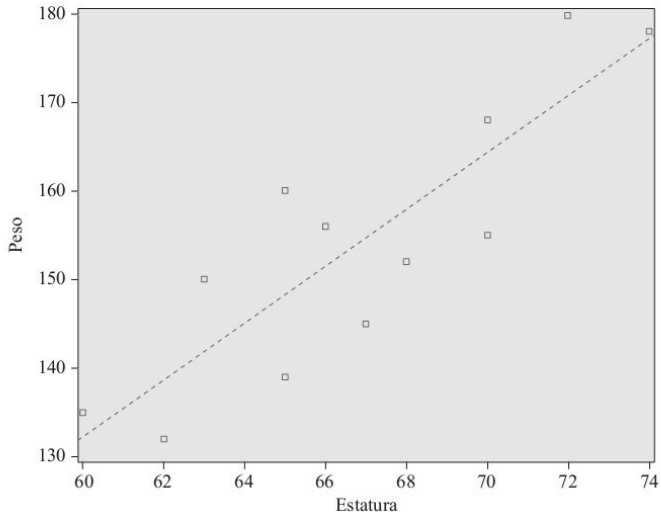


Figura 13-11 SAS, gráfica que presenta los puntos correspondientes a los datos de la tabla 13.5 y la recta de mínimos cuadrados.

13.13 *a*) Muestre que las dos rectas de mínimos cuadrados obtenidas en el problema 13.11 se intersecan en el punto $(\bar{X}, \bar{Y})$.

b) Estimar el valor de Y para $X = 12$.

c) Estimar el valor de X para $Y = 3$.

SOLUCIÓN

$$\bar{X} = \frac{\sum X}{N} = \frac{56}{8} = 7 \quad \bar{Y} = \frac{\sum Y}{N} = \frac{40}{8} = 5$$

Por lo tanto, el punto $(\bar{X}, \bar{Y})$, llamado el *centroide*, es (7, 5).

a) El punto (7, 5) se encuentra en la recta $Y = 0.545 + 0.636X$; o, más exactamente, $Y = \frac{6}{11} + \frac{7}{11}X$, ya que $5 = \frac{6}{11} + \frac{7}{11}(7)$. El punto (7, 5) se encuentra en la recta $X = -\frac{1}{2} + \frac{3}{2}Y$, ya que $7 = -\frac{1}{2} + \frac{3}{2}(5)$.

Otro método

Las ecuaciones de las dos rectas son $Y = \frac{6}{11} + \frac{7}{11}X$ y $X = -\frac{1}{2} + \frac{3}{2}Y$. Resolviendo simultáneamente estas dos ecuaciones se encuentra $X = 7$ y $Y = 5$. Por lo tanto, las rectas se intersecan en el punto (7, 5).

b) Sustituyendo $X = 12$ en la recta de regresión de Y (problema 13.11), $Y = 0.545 + 0.636(12) = 8.2$.

c) Sustituyendo $Y = 3$ en la recta de regresión de X (problema 13.11), $X = -0.50 + 1.50(3) = 4.0$.

13.14 Probar que una recta de mínimos cuadrados siempre pasa por el punto $(\bar{X}, \bar{Y})$.

SOLUCIÓN

Caso 1 (X es la variable independiente)

La ecuación de la recta de mínimos cuadrados es

$$Y = a_0 + a_1 X \tag{34}$$

Una de las ecuaciones normales de la recta de mínimos cuadrados es

$$\sum Y = a_0 N + a_1 \sum X \tag{35}$$

Dividiendo ambos lados de la ecuación (*35*) entre N se obtiene

$$\bar{Y} = a_0 + a_1 \bar{X} \tag{36}$$

Restando la ecuación (*36*) de la ecuación (*34*), la recta de mínimos cuadrados se puede escribir como

$$Y - \bar{Y} = a_1 (X - \bar{X}) \tag{37}$$

lo que muestra que la recta pasa a través del punto $(\bar{X}, \bar{Y})$.

Caso 2 (Y es la variable independiente)

Procediendo como en el caso 1, pero intercambiando X y Y y sustituyendo las constantes a_0 y a_1 por b_0 y b_1, respectivamente, se encuentra que la recta de mínimos cuadrados puede escribirse como

$$X - \bar{X} = b_1 (Y - \bar{Y}) \tag{38}$$

lo que indica que la recta pasa por el punto $(\bar{X}, \bar{Y})$.

Obsérvese que las rectas (*37*) y (*38*) no coinciden, sino que se intersecan en $(\bar{X}, \bar{Y})$.

13.15 *a*) Considerando X como la variable independiente, mostrar que la ecuación de la recta de mínimos cuadrados se puede escribir como

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \qquad \text{o bien} \qquad y = \left(\frac{\sum xY}{\sum x^2}\right)x$$

donde $x = X - \bar{X}$ y $y = Y - \bar{Y}$.

b) Si $\bar{X} = 0$, mostrar que la recta de mínimos cuadrados del inciso *a*) puede escribirse como

$$Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X$$

c) Dar la ecuación de la recta de mínimos cuadrados correspondiente a la del inciso *a*) en el caso en que Y sea la variable independiente.

d) Verificar que las rectas de los incisos *a*) y *c*) no son necesariamente iguales.

SOLUCIÓN

a) La ecuación (*37*) puede escribirse como $y = a_1 x$, donde $x = X - \bar{X}$ y $y = Y - \bar{Y}$. Además, resolviendo simultáneamente las ecuaciones normales (*18*) se tiene

$$\begin{aligned} a_1 &= \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = \frac{N\sum (x+\bar{X})(y+\bar{Y}) - [\sum (x+\bar{X}][\sum (y+\bar{Y})]}{N\sum (x+\bar{X})^2 - [\sum (x+\bar{X})]^2} \\ &= \frac{N\sum (xy + x\bar{Y} + \bar{X}y + \bar{X}\bar{Y}) - (\sum x + N\bar{X})(\sum y + N\bar{Y})}{N\sum (x^2 + 2x\bar{X} + \bar{X}^2) - (\sum x + N\bar{X})^2} \\ &= \frac{N\sum xy + N\bar{Y}\sum x + N\bar{X}\sum y + N^2\bar{X}\bar{Y} - (\sum x + N\bar{X})(\sum y + N\bar{Y})}{N\sum x^2 + 2N\bar{X}\sum x + N^2\bar{X}^2 - (\sum x + N\bar{X})^2} \end{aligned}$$

Pero $\sum x = \sum (X - \bar{X}) = 0$ y $\sum y = \sum (Y - \bar{Y}) = 0$; por lo que la fórmula anterior se simplifica a

$$a_1 = \frac{N\sum xy + N^2\bar{X}\bar{Y} - N^2\bar{X}\bar{Y}}{N\sum x^2 + N^2\bar{X}^2 - N^2\bar{X}^2} = \frac{\sum xy}{\sum x^2}$$

Lo que puede escribirse como

$$a_1 = \frac{\sum xy}{\sum x^2} = \frac{\sum x(Y - \bar{Y})}{\sum x^2} = \frac{\sum xY - \bar{Y}\sum x}{\sum x^2} = \frac{\sum xY}{\sum x^2}$$

Por lo tanto, la recta de mínimos cuadrados es $y = a_1 x$; es decir,

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \qquad \text{o bien} \qquad y = \left(\frac{\sum xY}{\sum x^2}\right)x$$

b) Si $\bar{X} = 0$, $x = X - \bar{X} = X$. Entonces, de acuerdo con la fórmula

$$y = \left(\frac{\sum xY}{\sum x^2}\right)$$

se tiene $$y = \left(\frac{\sum XY}{\sum X^2}\right)X \qquad \text{o bien} \qquad Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X$$

Otro método

Las ecuaciones normales de la recta de mínimos cuadrados $Y = a_0 + a_1 X$ son

$$\sum Y = a_0 N + a_1 \sum X \qquad \text{y} \qquad \sum XY = a_0 \sum X + a_1 \sum X^2$$

Si $\bar{X} = (\sum X)/N = 0$, entonces $\sum X = 0$ y las ecuaciones normales se transforman en

$$\sum Y = a_0 N \qquad \text{y} \qquad \sum XY = a_1 \sum X^2$$

de donde $$a_0 = \frac{\sum Y}{N} = \bar{Y} \qquad \text{y} \qquad a_1 = \frac{\sum XY}{\sum X^2}$$

Por lo tanto, la ecuación buscada de la recta de mínimos cuadrados es

$$Y = a_0 + a_1 X \qquad \text{o bien} \qquad Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X$$

c) Intercambiando X y Y o bien x y y, se puede demostrar como en el inciso *a*) que

$$x = \left(\frac{\sum xy}{\sum y^2}\right) y$$

d) De acuerdo con el inciso *a*), la recta de mínimos cuadrados es

$$y = \left(\frac{\sum xy}{\sum x^2}\right) x \tag{39}$$

De acuerdo con el inciso *c*), la recta de mínimos cuadrados es

$$x = \left(\frac{\sum xy}{\sum y^2}\right) y$$

o bien

$$y = \left(\frac{\sum y^2}{\sum xy}\right) x \tag{40}$$

Como en general

$$\frac{\sum xy}{\sum x^2} \neq \frac{\sum y^2}{\sum xy}$$

en general las rectas de mínimos cuadrados (*39*) y (*40*) son diferentes. Sin embargo, obsérvese que estas rectas se intersecan en $x = 0$ y $y = 0$ [es decir, en el punto $(\bar{X}, \bar{Y})$].

13.16 Si $X' = X + A$ y $Y' = Y + B$, donde A y B son constantes cualesquiera, probar que

$$a_1 = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} = \frac{N \sum X'Y' - (\sum X')(\sum Y')}{N \sum X'^2 - (\sum X')^2} = a_1'$$

SOLUCIÓN

$$x' = X' - \bar{X}' = (X + A) - (\bar{X} + A) = X - \bar{X} = x$$
$$y' = Y' - \bar{Y}' = (Y + B) - (\bar{Y} + B) = Y - \bar{Y} = y$$

Entonces

$$\frac{\sum xy}{\sum x^2} = \frac{\sum x'y'}{\sum x'^2}$$

y el resultado es consecuencia del problema 13.15. Un resultado similar es válido para b_1.

Este resultado es útil, pues permite simplificar los cálculos para obtener la recta de regresión sustrayendo a las variables X y Y constantes adecuadas (ver el segundo método del problema 13.17).

Nota: Este resultado no es válido si $X' = c_1X + A$ y $Y' = c_2Y + B$, a menos que $c_1 = c_2$.

13.17 Ajustar una recta de mínimos cuadrados a los datos del problema 13.10 empleando: *a*) X como la variable independiente y *b*) X como variable dependiente.

SOLUCIÓN

Primer método

a) De acuerdo con el problema 13.15*a*), la recta buscada es

$$y = \left(\frac{\sum xy}{\sum x^2}\right) x$$

donde $x = X - \bar{X}$ y $y = Y - \bar{Y}$. Los cálculos de las sumas se pueden organizar como se muestra en la tabla 13.7. De acuerdo con las dos primeras columnas $\bar{X} = 802/12 = 66.8$ y $\bar{Y} = 1\,850/12 = 154.2$. La última columna se incluyó para emplearla en el inciso *b*).

Tabla 13.7

Estatura X	Peso Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	xy	x^2	y^2
70	155	3.2	0.8	2.56	10.24	0.64
63	150	−3.8	−4.2	15.96	14.44	17.64
72	180	5.2	25.8	134.16	27.04	665.64
60	135	−6.8	−19.2	130.56	46.24	368.64
66	156	−0.8	1.8	−1.44	0.64	3.24
70	168	3.2	13.8	44.16	10.24	190.44
74	178	7.2	23.8	171.36	51.84	566.44
65	160	−1.8	5.8	−10.44	3.24	33.64
62	132	−4.8	−22.2	106.56	23.04	492.84
67	145	0.2	−9.2	−1.84	0.04	84.64
65	139	−1.8	−15.2	27.36	3.24	231.04
68	152	1.2	−2.2	−2.64	1.44	4.84
$\sum X = 802$ $\bar{X} = 66.8$	$\sum Y = 1\,850$ $\bar{Y} = 154.2$			$\sum xy = 616.32$	$\sum x^2 = 191.68$	$\sum y^2 = 2\,659.68$

La recta de mínimos cuadrados buscada es

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x = \frac{616.32}{191.68}\,x = 3.22x$$

o bien $Y - 154.2 = 3.22(X - 66.8)$, lo que puede escribirse como $Y = 3.22X - 60.9$. A esta ecuación se le conoce como la *recta de regresión de Y sobre X* y sirve para estimar valores de Y a partir de valores dados de X.

b) Si X es la variable dependiente, la recta buscada es

$$x = \left(\frac{\sum xy}{\sum y^2}\right)y = \frac{616.32}{2\,659.68}\,y = 0.232y$$

la cual se puede escribir como $X - 66.8 = 0.232(Y - 154.2)$, o bien $X = 31.0 + 0.232Y$. A esta ecuación se le conoce como la *recta de regresión de X sobre Y* y se utiliza para estimar X a partir de valores dados de Y.

Obsérvese que, si se desea, también se puede emplear el método del problema 13.11.

Segundo método

Empleando la fórmula del problema 13.16, de X y Y también se pueden sustraer cantidades adecuadas. Se sustraerá 65 a X y 150 a Y. Los cálculos se pueden organizar como en la tabla 13.7.

$$a_1 = \frac{N\sum X'Y' - (\sum X')(\sum Y')}{N\sum X'^2 - (\sum X')^2} = \frac{(12)(708) - (22)(50)}{(12)(232) - (22)^2} = 3.22$$

$$b_1 = \frac{N\sum X'Y' - (\sum Y')(\sum X')}{N\sum Y'^2 - (\sum Y')^2} = \frac{(12)(708) - (50)(22)}{(12)(2\,868) - (50)^2} = 0.232$$

Como $\bar{X} = 65 + 22/12 = 66.8$ y $\bar{Y} = 150 + 50/12 = 154.2$, las ecuaciones de regresión son $Y - 154.2 = 3.22(X - 66.8)$ y $X - 66.8 = 0.232(Y - 154.2)$; es decir, $Y = 3.22X - 60.9$ y $X = 0.232Y + 31.0$, en coincidencia con el primer método.

13.18 Resolver el problema 13.17 usando MINITAB. En un mismo conjunto de ejes, trazar la recta de regresión de pesos contra estaturas y la recta de regresión de estaturas contra pesos. Mostrar que el punto $(\bar{X}, \bar{Y})$ satisface ambas ecuaciones. Estas rectas se intersecan en $(\bar{X}, \bar{Y})$.

SOLUCIÓN

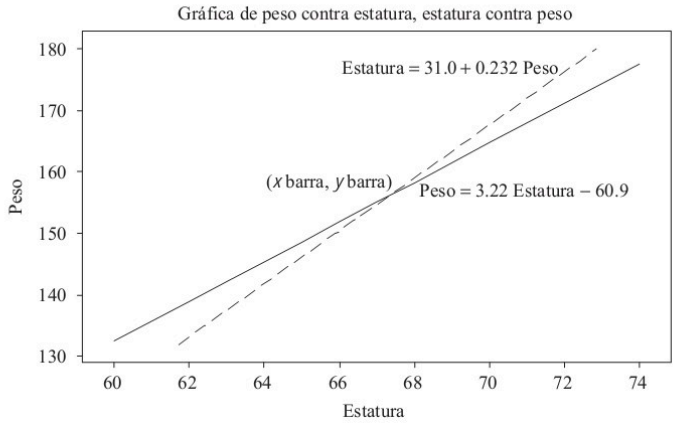


Figura 13-12 Tanto la recta de regresión de estatura contra peso como la recta de regresión del peso contra estatura pasan a través del punto (*x* barra, *y* barra).

$(\bar{X}, \bar{Y})$ es lo mismo que (x barra, y barra) y es igual a (66.83, 154.17). Obsérvese que peso $= 3.22(\mathbf{66.83}) - 60.9 =$ **154.17** y estatura $= 31.0 + 0.232(\mathbf{154.17}) =$ **66.83**. Por lo tanto, ambas rectas pasan a través de (x barra, y barra).

Tabla 13.8

X'	Y'	X'^2	$X'Y'$	Y'^2
5	5	25	25	25
−2	0	4	0	0
7	30	49	210	900
−5	−15	25	75	225
1	6	1	6	36
5	18	25	90	324
9	28	81	252	784
0	10	0	0	100
−3	−18	9	54	324
2	−5	4	−10	25
0	−11	0	0	121
3	2	9	6	4
$\sum X' = 22$	$\sum Y' = 50$	$\sum X'^2 = 232$	$\sum X'Y' = 708$	$\sum Y'^2 = 2\,868$

APLICACIONES PARA SERIES DE TIEMPO

13.19 En la tabla 13.9 se presentan, en millones de dólares, las exportaciones agrícolas de Estados Unidos. Usar MINITAB para hacer lo siguiente:

Tabla 13.9

Año	2000	2001	2002	2003	2004	2005
Valor total	51 246	53 659	53 115	59 364	61 383	62 958
Código del año	1	2	3	4	5	6

Fuente: The 2007 Statistical Abstract.

a) Graficar los datos y mostrar la recta de regresión de mínimos cuadrados.
b) Encontrar y graficar la recta de tendencia de los datos.
c) Dar los *valores ajustados* y los *residuales* empleando los códigos de los años.
d) Estimar el valor de las exportaciones agrícolas en 2006.

SOLUCIÓN

a) En la figura 13-13*a*) se muestran los datos y la recta de regresión. La gráfica que se muestra en la figura 13-13*a*) se obtiene empleando la secuencia **Stat** → **Regresión** → **Fitted line plot**.

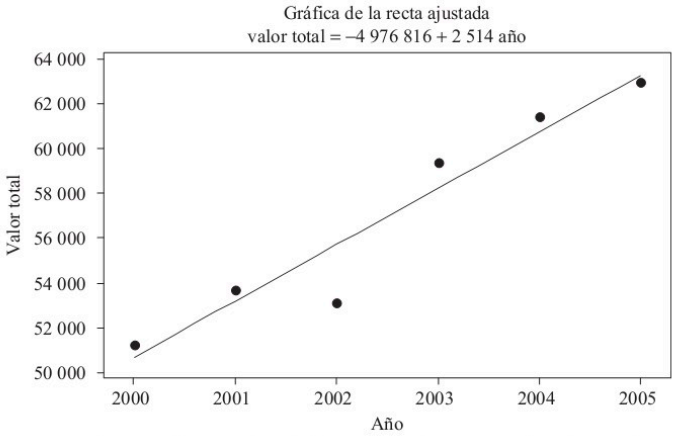


Figura 13-13 ***a*) Recta de regresión de las exportaciones agrícolas de Estados Unidos dadas en millones de dólares.**

b) La gráfica que se muestra en la figura 13.13*b*) se obtiene empleando la secuencia **Stat** → **Time series** → **Trend Análisis**. Ésta es una manera diferente de ver los mismos datos. Tal vez sea un poco más fácil emplear los números índice (códigos de los años) en vez de los años.

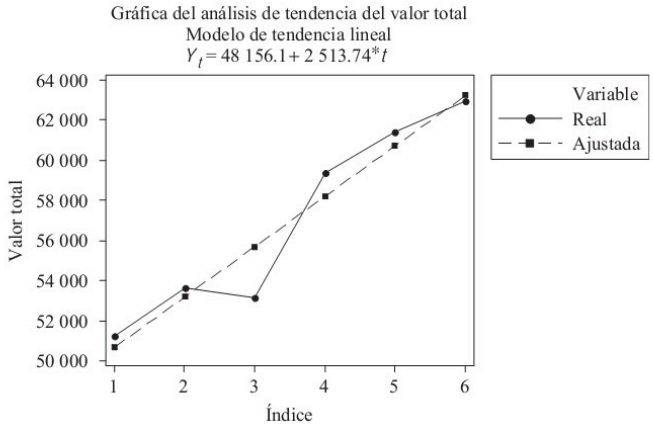


Figura 13-13 ***b*) Recta de tendencia de las exportaciones agrícolas de Estados Unidos, dadas en millones de dólares.**

c) En la tabla 13.10 se dan los valores ajustados y los residuales de los datos que se presentan en la tabla 13.9; se emplean años codificados.

Tabla 13.10

Año codificado	Valor total	Valor ajustado	Residual
1	51 246	50 669.8	576.19
2	53 659	53 183.6	475.45
3	53 115	55 697.3	−2 582.30
4	59 364	58 211.0	1 152.96
5	61 383	60 724.8	658.22
6	62 958	63 238.5	−280.52

d) Empleando el año codificado, el valor estimado es $Y_t = 48\,156.1 + 2\,513.74(7) = 65\,752.3$.

13.20 En la tabla 13.11 se presenta el poder de compra del dólar, medido a través de los precios al consumidor, de acuerdo con lo informado por la Oficina de Estadísticas Laborales de Estados Unidos.

Tabla 13.11

Año	2000	2001	2002	2003	2004	2005
Precios al consumidor	0.581	0.565	0.556	0.544	0.530	0.512

Fuente: U. S. Bureau of Labor Statistics, Survey of Current Business.

a) Graficar los datos y obtener la recta de tendencia usando MINITAB.
b) Encontrar, a mano, la ecuación de la línea de tendencia.
c) Estimar el precio al consumidor del 2008 suponiendo que la tendencia continúe tres años más.

SOLUCIÓN

a) En la figura 13-14, la línea continua es la gráfica de los datos de la tabla 13.11 y la línea punteada es la gráfica de la recta de mínimos cuadrados.

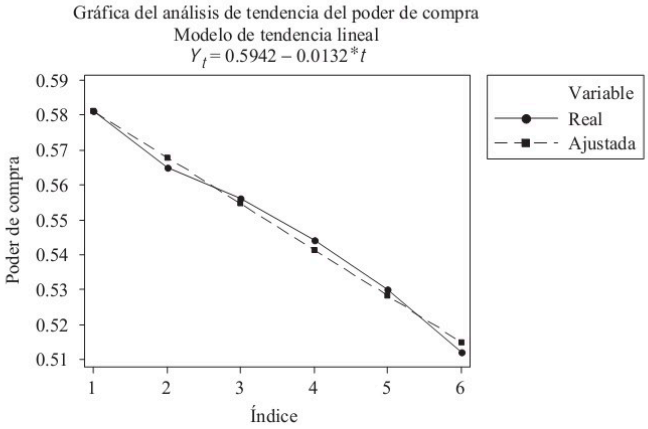


Figura 13-14 Línea de tendencia del poder de compra.

b) En la tabla 13.12 se presentan los cálculos para hallar, a mano, la línea de tendencia. La ecuación es

$$y = \frac{\sum xy}{\sum x^2} x$$

donde $x = X - \bar{X}$ y $y = Y - \bar{Y}$; por lo que esta ecuación se puede escribir como $Y - 0.548 = -0.0132(X - 3.5)$ o $Y = -0.0132X + 0.5942$. Como se ilustra con este problema, el trabajo que se ahorra empleando algún software para estadística es enorme.

Tabla 13.12

Año	X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy
2000	1	0.581	−2.5	0.033	6.25	−0.0825
2001	2	0.565	−1.5	0.017	2.25	−0.0255
2002	3	0.556	−0.5	0.008	0.25	−0.004
2003	4	0.544	0.5	−0.004	0.25	−0.002
2004	5	0.530	1.5	−0.018	2.25	−0.027
2005	6	0.512	2.5	−0.036	6.25	−0.09
	$\Sigma X = 21$	$\Sigma Y = 3.288$			Σx^2	Σxy
	$\bar{X} = 3.5$	$\bar{Y} = 0.548$			17.5	−0.231

c) El precio al consumidor estimado del 2008 se obtiene sustituyendo en la ecuación de la línea tendencia $X = 9$. El precio al consumidor estimado es $0.5942 - 0.0132(9) = 0.475$.

ECUACIONES NO LINEALES REDUCIBLES A LA FORMA LINEAL

13.21 En la tabla 13.13 se dan los valores experimentales de la presión P de una masa dada de gas correspondientes a diversos valores del volumen V. De acuerdo con los principios de la termodinámica, entre estas variables existe una relación de la fórmula $PV^{\gamma} = C$, donde γ y C son constantes.

a) Encontrar los valores de γ y de C.

b) Escribir la ecuación que relaciona P y V.

Tabla 13.13

Volumen V en pulgadas cúbicas (in^3)	54.3	61.8	72.4	88.7	118.6	194.0
Presión P en libras por pulgada cuadrada (lb/in^2)	61.2	49.2	37.6	28.4	19.2	10.1

c) Estimar P para $V = 100.0$ (lb/in^2).

SOLUCIÓN

Como $PV^{\gamma} = C$, se tiene

$$\log P + \gamma \log V = \log C \qquad \text{o bien} \qquad \log P = \log C - \gamma \log V$$

Haciendo $\log V = X$ y $\log P = Y$, la última ecuación puede escribirse como

$$Y = a_0 + a_1 X \tag{41}$$

donde $a_0 = \log C$ y $a_1 = -\gamma$.

En la tabla 13.14 se dan los valores de $X = \log V$ y de $Y = \log P$, correspondientes a los valores de V y P dados en la tabla 13.13, y se indican también los cálculos para obtener la recta (*41*) de mínimos cuadrados. Las ecuaciones normales correspondientes a la recta (*41*) de mínimos cuadrados son

$$\sum Y = a_0 N + a_1 \sum X \qquad \text{y} \qquad \sum XY = a_0 \sum X + a_1 \sum X^2$$

de donde

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2} = 4.20 \qquad a_1 = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} = -1.40$$

Por lo tanto, $Y = 4.20 - 1.40X$.

a) Como $a_0 = 4.20 = \log C$ y $a_1 = -1.40 = -\gamma$, $C = 1.60 \times 10^4$ y $\gamma = 1.40$.

b) La ecuación que se busca en términos de P y V se puede escribir como $PV^{1.40} = 16\,000$.

c) Para $V = 100$, $X = \log V = 2$ y $Y = \log P = 4.20 - 1.40(2) = 1.40$. Entonces $P = \text{antilog } 1.40 = 25.1 \text{ lb/in}^2$.

Tabla 13.14

$X = \log V$	$Y = \log P$	X^2	XY
1.7348	1.7868	3.0095	3.0997
1.7910	1.6946	3.2077	3.0350
1.8597	1.5752	3.4585	2.9294
1.9479	1.4533	3.7943	2.8309
2.0741	1.2833	4.3019	2.6617
2.2878	1.0043	5.2340	2.2976
$\sum X = 11.6953$	$\sum Y = 8.7975$	$\sum X^2 = 23.0059$	$\sum XY = 16.8543$

13.22 Usar MINITAB para resolver el problema 13.21.

SOLUCIÓN

Las transformaciones $X = \log_t(V)$ y $Y = \log_t(P)$ convierten el problema en un problema de ajuste lineal. Para encontrar los logaritmos comunes del volumen y de la presión se emplea la calculadora de MINITAB. En las columnas `C1` a `C4` de la hoja de cálculo de MINITAB se tendrá:

V	P	Log10 V	Log10 P
54.3	61.2	1.73480	1.78675
61.8	49.2	1.79099	1.69197
72.4	37.6	1.85974	1.57519
88.7	28.4	1.94792	1.45332
118.6	19.2	2.07408	1.28330
194.0	10.1	2.28780	1.00432

El ajuste por mínimos cuadrados da: $\log_{10}(P) = 4.199 - 1.402 \log_{10}(V)$. Ver la figura 13-15. $a_0 = \log C$ y $a_1 = -\gamma$. Sacando antilogaritmos se obtiene $C = 10^{a_0}$ y $\gamma = -a_1$ o $C = 15\,812$ y $\gamma = 1.402$. La ecuación no lineal es $PV^{1.402} = 15\,812$.

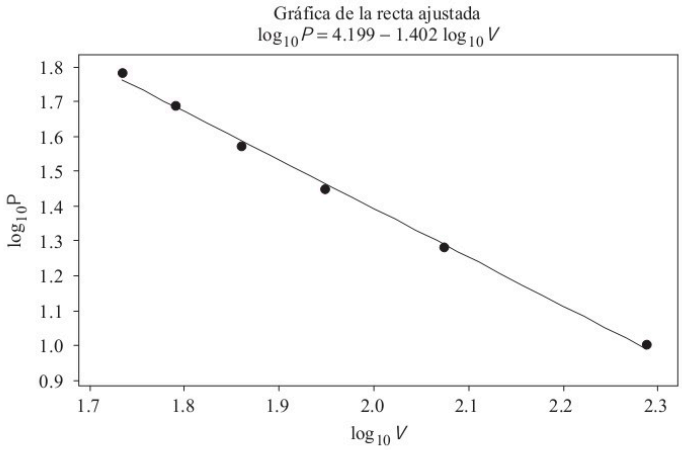


Figura 13-15 Reducción de una ecuación no lineal a la forma lineal.

13.23 En la tabla 13.15 se da, en millones, la población de Estados Unidos desde 1960 hasta 2005. A estos datos, ajustar una recta y una parábola y analizar los dos ajustes. Usar ambos modelos para predecir la población que tendrá Estados Unidos en 2010.

Tabla 13.15

Año	1960	1965	1970	1975	1980	1985	1990	1995	2000	2005
Población	181	194	205	216	228	238	250	267	282	297

Fuente: U.S. Bureau of Census.

SOLUCIÓN

A continuación se presenta parte de los resultados que da MINITAB para la recta de mínimos cuadrados y para la parábola de mínimos cuadrados.

Año	Población	*x*	*x*cuadrada
1960	181	1	1
1965	194	2	4
1970	205	3	9
1975	216	4	16
1980	228	5	25
1985	238	6	36
1990	250	7	49
1995	267	8	64
2000	282	9	81
2005	297	10	100

El modelo para la recta es el siguiente:
La ecuación de regresión es

Población = 166 + 12.6 *x*

El modelo cuadrático es el siguiente:
La ecuación de regresión es

Población = 174 + 9.3 *x* – 0.326 x^2

En la tabla 13.16 se dan los valores ajustados y los residuales del ajuste a los datos mediante la recta.

Tabla 13.16

Año	Población	Valor ajustado	Residual
1960	181	179.018	1.98182
1965	194	191.636	2.36364
1970	205	204.255	0.74545
1975	216	216.873	−0.87273
1980	228	229.491	−1.49091
1985	238	242.109	−4.10909
1990	250	254.727	−4.72727
1995	267	267.345	−0.34545
2000	282	279.964	2.03636
2005	297	292.582	4.41818

En la tabla 13.17 se dan los valores ajustados y los residuales correspondientes al ajuste parabólico a los datos. La suma de los cuadrados de los residuales en el caso de la recta es 76.073 y la suma de los cuadrados de los residuales en el caso de la parábola es 20.042. Parece que, en general, la parábola se ajusta mejor que la recta a estos datos.

Tabla 13.17

Año	Población	Valor ajustado	Residual
1960	181	182.927	−1.92727
1965	194	192.939	1.06061
1970	205	203.603	1.39697
1975	216	214.918	1.08182
1980	228	226.885	1.11515
1985	238	239.503	−1.50303
1990	250	252.773	−2.77273
1995	267	266.694	0.30606
2000	282	281.267	0.73333
2005	297	296.491	0.50909

Para predecir cuál será la población en el año 2010, obsérvese que el código para 2010 es 11. El valor que se obtiene con el modelo de la recta es población $= 166 + 12.6x = 166 + 138.6 = 304.6$ millones y con el modelo de la parábola es población $= 174 + 9.03x + 0.326x^2 = 174 + 99.33 + 39.446 = 312.776$.

PROBLEMAS SUPLEMENTARIOS

LÍNEAS RECTAS

13.24 Si $3X + 2Y = 18$, encontrar: *a*) el valor de X para $Y = 3$, *b*) el valor de Y para $X = 2$, *c*) el valor de X para $Y = -5$, *d*) el valor de Y para $X = -1$, *e*) la intersección con el eje X, y *f*) la intersección con el eje Y.

13.25 En un mismo conjunto de ejes, trazar la gráfica de las ecuaciones: *a*) $Y = 3X - 5$ y *b*) $X + 2Y = 4$. ¿En qué punto se intersecan?

13.26 *a*) Encontrar la ecuación de la recta que pasa por los puntos $(3, -2)$ y $(-1, 6)$.
b) Determinar las intersecciones de la recta del inciso *a*) con el eje X y con el eje Y.
c) Encontrar el valor de Y que corresponde a $X = 3$ y a $X = 5$.
d) A partir de la gráfica, verificar sus respuestas a los incisos *a*), *b*) y *c*).

13.27 Encontrar la ecuación de la recta cuya pendiente es $\frac{2}{3}$ y cuya intersección con el eje Y es -3.

13.28 *a*) Encontrar la pendiente y la intersección con el eje Y de la recta cuya ecuación es $3X - 5Y = 20$.
b) ¿Cuál es la ecuación de la recta paralela a la recta del inciso *a*) y qué pasa por el punto $(2, -1)$?

13.29 Encontrar: *a*) la pendiente, *b*) la intersección con el eje Y y *c*) la ecuación de la recta que pasa por los puntos (5, 4) y (2, 8).

13.30 Encontrar la ecuación de la recta cuyas intersecciones con los ejes X y Y son 3 y -5, respectivamente.

13.31 La temperatura de 100 grados Celsius (°C) corresponde a 212 grados Fahrenheit (°F), en tanto que la temperatura de 0°C corresponde a 32°F. Suponiendo que exista una relación lineal entre temperaturas Celsius y temperaturas Fahrenheit, encontrar: *a*) la ecuación que relaciona temperaturas Celsius y temperaturas Fahrenheit, *b*) la temperatura Fahrenheit que corresponde a 80°C y *c*) la temperatura Celsius que corresponde 68°F.

LA RECTA DE MÍNIMOS CUADRADOS

13.32 Ajustar una recta de mínimos cuadrados a los datos de la tabla 13.18 usando: *a*) X como la variable independiente y *b*) X como la variable dependiente. Graficar los datos de estas rectas de mínimos cuadrados en un mismo eje de coordenadas.

Tabla 13.18

X	3	5	6	8	9	11
Y	2	3	4	6	5	8

13.33 Dados los datos del problema 13.32, hallar: *a*) el valor de Y para $X = 12$ y *b*) el valor de X para $Y = 7$.

13.34 *a*) Empleando el método a mano, obtener una ecuación de la recta que se ajuste a los datos del problema 13.32.
b) Empleando el resultado del inciso *a*), resolver el problema 13.33.

13.35 En la tabla 13.19 se muestran las calificaciones finales de álgebra y de física de diez estudiantes, tomados en forma aleatoria de un grupo grande.

a) Graficar los datos.
b) Encontrar la recta de mínimos cuadrados que se ajusta a los datos, usando X como la variable independiente.

c) Encontrar la recta de mínimos cuadrados que se ajusta a los datos, usando Y como la variable independiente.
d) Si la calificación de un estudiante en álgebra es 75, ¿cuál es la calificación que se espera que obtenga en física?
e) Si la calificación de un estudiante en física es 95, ¿cuál es la calificación que se espera que obtenga en álgebra?

Tabla 13.19

Álgebra (X)	75	80	93	65	87	71	98	68	84	77
Física (Y)	82	78	86	72	91	80	95	72	89	74

13.36 En la tabla 13.20 se muestra la tasa de nacimiento por cada mil personas desde 1998 hasta 2004.

a) Graficar estos datos.
b) Hallar la recta de mínimos cuadrados que se ajusta a estos datos. Asignar a los años 1998 a 2004 los números 1 a 7.
c) Calcular los valores de tendencia (valores ajustados) y los residuales.
d) Indicar cuál será la tasa de nacimiento en 2010, suponiendo que la tendencia actual continúa.

Tabla 13.20

Año	1998	1999	2000	2001	2002	2003	2004
Tasa de nacimientos por cada 1 000	14.3	14.2	14.4	14.1	13.9	14.1	14.0

Fuente: U.S. Nacional Center for Health Statistics, Vital Statistics of the United Status, annual; Nacional Vital Statistics Reports y datos inéditos.

13.37 En la tabla 13.21 se presenta, en miles, la población de Estados Unidos de 85 o más años, desde 1999 hasta 2005.

a) Graficar estos datos.
b) Encontrar la recta de mínimos cuadrados que se ajusta a estos datos. Asignar a los años 1999 a 2005 los números 1 a 7.
c) Calcular los valores de tendencia (valores ajustados) y los residuales.
d) Suponiendo que la tendencia actual continúe, indicar cuál será el número de personas de 85 años o más en el 2010.

Tabla 13.21

Año	1999	2000	2001	2002	2003	2004	2005
85 o más	4 154	4 240	4 418	4 547	4 716	4 867	5 096

Fuente: U.S. Bureau of Census.

CURVAS DE MÍNIMOS CUADRADOS

13.38 Ajustar una parábola de mínimos cuadrados, $Y = a_0 + a_1X + a_2X^2$, a los datos de la tabla 13.22.

Tabla 13.22

X	0	1	2	3	4	5	6
Y	2.4	2.1	3.2	5.6	9.3	14.6	21.9

13.39 El tiempo requerido para llevar un automóvil al alto total a partir de que se percibe un peligro es el tiempo de reacción (el tiempo entre el reconocimiento del peligro y la aplicación del freno) más el tiempo de frenado (el tiempo necesario para que el automóvil se detenga después de la aplicación del freno). En la tabla 13.23 se da la distancia de frenado D (en pies, ft) de un automóvil que va a una velocidad V (en millas por hora, mi/h).

a) Graficar D contra V.

b) Ajustar a estos datos una parábola de mínimos cuadrados de la forma $D = a_0 + a_1V + a_2V^2$.

c) Estimar D para $V = 45$ mi/h y 80 mi/h.

Tabla 13.23

Velocidad V (mi/h)	20	30	40	50	60	70
Distancia de frenado D (ft)	54	90	138	206	292	396

13.40 En la tabla 13.24 se presenta, en millones, la población de hombres y de mujeres en Estados Unidos, desde 1940 hasta 2005. Se presentan también los números dados como códigos a los años y la diferencia de hombres menos mujeres.

a) Graficar los datos y la recta de mejor ajuste por mínimos cuadrados.

b) Graficar los datos y el mejor ajuste cuadrático por mínimos cuadrados.

c) Graficar los datos y el mejor ajuste cúbico por mínimos cuadrados.

d) Con cada uno de los tres modelos, dar el valor ajustado y los residuales, así como la suma de los cuadrados de los residuales.

e) Emplear cada uno de los tres modelos para predecir la población que habrá en el año 2010.

Tabla 13.24

Año	1940	1950	1960	1970	1980	1990	2000	2005
Código	0	1	2	3	4	5	6	6.5
Hombres	66.1	75.2	88.3	98.9	110.1	121.2	138.1	146.0
Mujeres	65.6	76.1	91.0	104.3	116.5	127.5	143.4	150.4
Diferencia	0.5	−0.9	−2.7	−5.4	−6.4	−6.3	−5.3	−4.4

Fuente: U.S. Bureau of Census.

13.41 Resolver el problema 13.40 empleando, en lugar de las diferencias, la proporción entre mujeres y hombres.

13.42 Resolver el problema 13.40 ajustando una parábola de mínimos cuadrados a las diferencias.

13.43 En la tabla 13.25 se presenta la cuenta bacteriana Y, por unidad de volumen en un cultivo, después de X horas.

Tabla 13.25

Número de horas (X)	0	1	2	3	4	5	6
Cuenta bacteriana por unidad de volumen (Y)	32	47	65	92	132	190	275

a) Graficar los datos en papel semilogarítmico usando la escala logarítmica para Y y la escala aritmética para X.

b) Ajustar a los datos una curva de mínimos cuadrados de la forma $Y = ab^x$ y explicar por qué esta ecuación dará buenos resultados.

c) Comparar los valores de Y que se obtienen con esta ecuación con los valores reales.

d) Estimar el valor de Y para $X = 7$.

13.44 En el problema 13.43 mostrar cómo usar una gráfica en papel semilogarítmico para obtener la ecuación buscada sin emplear el método de mínimos cuadrados.

TEORÍA DE LA CORRELACIÓN

14

CORRELACIÓN Y REGRESIÓN

En el capítulo 13 se consideró el problema de la *regresión*, o *estimación* de una variable (la variable dependiente) a partir de una o más variables (las variables independientes). En este capítulo se hará referencia a un problema relacionado con el de la *correlación* o grado de relación entre las variables, en el que se busca determinar *qué tan bien* una ecuación lineal, o de otro tipo, describe o explica la relación entre las variables.

Si todos los valores de las variables satisfacen con exactitud una ecuación, se dice que las variables están en *perfecta correlación* o que hay una *correlación perfecta* entre ellas. Así, las circunferencias C y los radios r de todos los círculos están perfectamente correlacionados, ya que $C = 2\pi r$. Cuando se lanzan 100 veces dos dados en forma simultánea entre los puntos que aparecen en cada uno de ellos no hay relación alguna (a menos que estén cargados); es decir, no están *correlacionados*. Sin embargo, variables como el peso y la estatura de una persona muestran *cierta* correlación.

Cuando intervienen sólo dos variables se habla de *correlación simple* y de *regresión simple*. Cuando intervienen más de dos variables, se habla de *correlación múltiple* y de *regresión múltiple*. En este capítulo sólo se considerará la correlación simple. En el capítulo 15 se consideran la correlación y la regresión múltiples.

CORRELACIÓN LINEAL

Si X y Y son las dos variables en consideración, un *diagrama de dispersión* sirve para mostrar la localización de los puntos (X, Y) en un sistema de coordenadas rectangulares. Si en este diagrama de dispersión todos los puntos parecen encontrarse cerca de una línea recta, como en las figuras 14-1*a*) y 14-1*b*), a la correlación se le llama *lineal*. En estos casos, como se vio en el capítulo 13, una ecuación lineal es lo apropiado con el propósito de regresión (o estimación).

Si Y tiende a aumentar a medida que X aumenta, como en la figura 14-1*a*), se dice que la correlación es una *correlación positiva* o *directa*. Si Y tiende a disminuir a medida que X aumenta, como en la figura 14-1*b*), se dice que es una *correlación negativa* o *inversa*.

Si todos los puntos parecen encontrarse en una curva, esta correspondencia se llama *no lineal*, y según se vio en el capítulo 13, lo apropiado para la regresión es una ecuación no lineal. Es claro que la correlación no lineal puede ser algunas veces positiva y otras veces negativa.

Si no parece haber relación entre las variables, como en la figura 14-1*c*), se dice que *no hay relación* entre ellas (es decir, están *descorrelacionadas*).

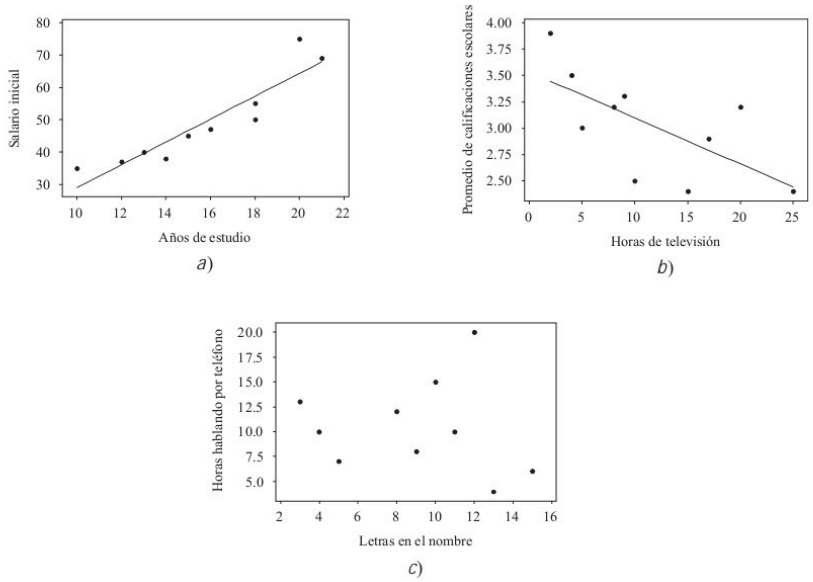


Figura 14-1 Ejemplos de correlación positiva, correlación negativa y ninguna correlación. *a*) El salario inicial y los años de estudio se correlacionan en forma positiva; *b*) el promedio de las calificaciones escolares y las horas que se pasa viendo la televisión se correlacionan negativamente; *c*) entre la cantidad de horas que se habla por teléfono y el número de letras que tiene el nombre de una persona no hay correlación.

MEDIDAS DE LA CORRELACIÓN

Mediante observación directa se puede determinar de manera *cualitativa* que también una recta o una curva describe la relación entre las variables. Por ejemplo, se ve que una línea recta es mucho más útil para describir la relación entre X y Y en el caso de los datos de la figura 14-1*a*) que en el caso de los datos de la figura 14-1*b*), debido a que en la figura 14-1*a*) hay menos dispersión con relación a la recta.

Para ocuparse de manera *cuantitativa* del problema de la dispersión de los datos muestrales respecto a una línea o a una curva, es necesario encontrar una *medida de la correlación*.

LAS RECTAS DE REGRESIÓN DE MÍNIMOS CUADRADOS

Primero se considerará el problema de qué tan bien una línea recta explica la relación entre dos variables. Para esto, se necesitarán las ecuaciones de las rectas de regresión por mínimos cuadrados obtenidas en el capítulo 13. Como se ha visto, la recta de regresión por mínimos cuadrados de Y sobre X es

$$Y = a_0 + a_1 X \tag{1}$$

donde a_0 y a_1 se obtienen de las ecuaciones normales

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 \end{aligned} \tag{2}$$

que dan

$$\begin{aligned} a_0 &= \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2} \\ a_1 &= \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} \end{aligned} \tag{3}$$

De igual manera, la recta de regresión de X sobre Y es

$$X = b_0 + b_1 Y \tag{4}$$

donde b_0 y b_1 se obtienen de las ecuaciones normales

$$\begin{aligned} \sum X &= b_0 N + b_1 \sum Y \\ \sum XY &= b_0 \sum X + b_1 \sum Y^2 \end{aligned} \tag{5}$$

que dan

$$\begin{aligned} b_0 &= \frac{(\sum X)(\sum Y^2) - (\sum Y)(\sum XY)}{N \sum Y^2 - (\sum Y)^2} \\ b_1 &= \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum Y^2 - (\sum Y)^2} \end{aligned} \tag{6}$$

Las ecuaciones (*1*) y (*4*) pueden expresarse, respectivamente, como

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \qquad \text{y} \qquad x = \left(\frac{\sum xy}{\sum y^2}\right)y \tag{7}$$

donde $x = X - \bar{X}$ y $y = Y - \bar{Y}$.

Las ecuaciones de regresión son idénticas si y sólo si todos los puntos del diagrama de dispersión se encuentran en una recta. En tales casos, existe una *correlación lineal perfecta* entre X y Y.

EL ERROR ESTÁNDAR DE ESTIMACIÓN

Si Y_{est} es el valor estimado para Y, empleando la ecuación (*1*), para un valor dado de X, una medida de la dispersión respecto a la recta de regresión de Y sobre X es la cantidad

$$s_{Y.X} = \sqrt{\frac{\sum (Y - Y_{est})^2}{N}} \tag{8}$$

a la que se le llama *error estándar de estimación de Y sobre X*.

Empleando la recta de regresión (*4*), el error estándar de estimación análogo, de *X* sobre *Y*, es

$$s_{X.Y} = \sqrt{\frac{\sum (X - X_{\text{est}})^2}{N}} \tag{9}$$

En general, $s_{Y.X} \neq s_{X.Y}$.

La ecuación (*8*) también puede expresarse en la forma

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N} \tag{10}$$

que puede ser más apropiada para hacer los cálculos (ver problema 14.3). Para la ecuación (*9*) existe una expresión similar.

El error estándar de estimación tiene propiedades análogas a la desviación estándar. Por ejemplo, si se trazan rectas paralelas a la recta de regresión de *Y* sobre *X* a las distancias verticales $s_{Y.X}$, $2s_{Y.X}$ y $3s_{Y.X}$, se hallará, si *N* es suficientemente grande, que entre estas rectas se encuentra 68%, 95% y 99.7% de los puntos muestrales, respectivamente.

Así como la desviación estándar modificada, que es

$$\hat{s} = \sqrt{\frac{N}{N-1}}\, s$$

se emplea para muestras pequeñas, también el error estándar de estimación modificado está dado por

$$\hat{s}_{Y.X} = \sqrt{\frac{N}{N-2}}\, s_{Y.X}$$

A esto se debe que algunos especialistas en estadística prefieran definir las ecuaciones (*8*) y (*9*) empleando $N - 2$ en el denominador en lugar de *N*.

VARIACIÓN EXPLICADA Y NO EXPLICADA

La *variación total de Y* se define como $\sum (Y - \bar{Y})^2$; es decir, la suma de los cuadrados de las desviaciones de *Y* respecto a la media $\bar{Y}$. Como se muestra en el problema 14.7, esta expresión se puede expresar como

$$\sum (Y - \bar{Y})^2 = \sum (Y - Y_{\text{est}})^2 + \sum (Y_{\text{est}} - \bar{Y})^2 \tag{11}$$

En la ecuación (*11*), al primer término del lado derecho se le llama *variación no explicada*, en tanto que al segundo término se le llama *variación explicada*; se les llama así debido a que las desviaciones $Y_{\text{est}} - \bar{Y}$ tienen un patrón definido; en cambio, las desviaciones $Y - Y_{\text{est}}$ son aleatorias o impredecibles. Para la variable *X* existe una fórmula similar.

COEFICIENTE DE CORRELACIÓN

Al cociente de la variación explicada entre la variación total se le llama *coeficiente de determinación*. Si hay cero variación explicada (es decir, si la variación total es sólo variación no explicada), este cociente es 0. Si hay 0 variación no explicada (es decir, si la variación total es sólo variación explicada), este cociente es 1. En los demás casos, este cociente se encuentra entre 0 y 1; como siempre es no negativo, se denota r^2. A la cantidad *r* se le llama *coeficiente de correlación*; está dado por

$$r = \pm\sqrt{\frac{\text{variación explicada}}{\text{variación total}}} = \pm\sqrt{\frac{\sum (Y_{\text{est}} - \bar{Y})^2}{\sum (Y - \bar{Y})^2}} \tag{12}$$

y varía entre -1 y $+1$. Los signos $+$ y $-$ se usan para correlación lineal positiva y correlación lineal negativa, respectivamente. Obsérvese que r es una cantidad adimensional; es decir, no depende de las unidades que se empleen.

Utilizando las ecuaciones (*8*) y (*11*) y el hecho de que la desviación estándar de Y es

$$s_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}} \tag{13}$$

se encuentra que la ecuación (*12*) puede expresarse, sin hacer caso del signo, como

$$r = \sqrt{1 - \frac{s_{Y.X}^2}{s_Y^2}} \qquad \text{o bien} \qquad s_{Y.X} = s_Y \sqrt{1 - r^2} \tag{14}$$

Si se intercambian X y Y se obtienen ecuaciones similares.

En el caso de la correlación lineal, la cantidad r es la misma, ya sea que se considere a X o a Y como la variable independiente. Por lo tanto r es una muy buena medida de la correlación lineal entre dos variables.

OBSERVACIONES ACERCA DEL COEFICIENTE DE CORRELACIÓN

Las definiciones del coeficiente de correlación dadas en las ecuaciones (*12*) y (*14*) son muy generales y pueden emplearse tanto para relaciones no lineales como para relaciones lineales; la única diferencia es que Y_{est} se calcula a partir de una ecuación de regresión no lineal y no a partir de una ecuación de regresión lineal, y que los signos $+$ y $-$ se omiten. En estos casos la ecuación (*8*), que define el error estándar de estimación, es perfectamente general. Sin embargo, la ecuación (*10*) que se emplea únicamente para regresión lineal, debe ser modificada. Si, por ejemplo, la ecuación de estimación es

$$Y = a_0 + a_1 X + a_2 X^2 + \cdots + a_{n-1} X^{n-1} \tag{15}$$

la ecuación (*10*) se reemplaza por

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY - \cdots - a_{n-1} \sum X^{n-1} Y}{N} \tag{16}$$

En este caso, el *error estándar de estimación modificado* (antes visto en este capítulo) es

$$\hat{s}_{Y.X} = \sqrt{\frac{N}{N-n}}\, s_{Y.X}$$

en donde a la cantidad $N - n$ se le conoce como *número de grados de libertad*.

Hay que subrayar que en todos los casos, el valor calculado para r mide el grado de relación respecto al tipo de ecuación que se emplee. Así, si se utiliza una ecuación lineal y con la ecuación (*12*) o (*14*) dan un valor de r cercano a cero, esto significa que entre las variables casi no hay *correlación lineal*. Pero esto no significa que no haya correlación alguna, pues entre estas variables puede haber una fuerte *correlación no lineal*. En otras palabras, el coeficiente de correlación mide la bondad de ajuste entre: 1) la ecuación empleada y 2) los datos. A menos que se especifique otra cosa, el término *coeficiente de correlación* se emplea con el significado de *coeficiente de correlación lineal*.

Hay que hacer notar también que un coeficiente de correlación elevado (es decir, cercano a 1 o a -1) no necesariamente indica que haya dependencia directa entre las variables. Así, por ejemplo, puede haber correlación elevada entre la cantidad de libros publicados anualmente y cantidades número de tormentas eléctricas por año. A los ejemplos de este tipo o se le conoce como *correlaciones sin sentido* o *espurias*.

FÓRMULA PRODUCTO-MOMENTO PARA EL COEFICIENTE DE CORRELACIÓN LINEAL

Si se supone que entre dos variables existe una relación lineal, la ecuación (*12*) se convierte en

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} \tag{17}$$

donde $x = X - \bar{X}$ y $y = Y - \bar{Y}$ (ver el problema 14.10). Esta fórmula, que automáticamente da el signo adecuado de *r* se conoce como *fórmula del producto-momento* y permite ver claramente la simetría entre *X* y *Y*.

Si se escribe

$$s_{XY} = \frac{\sum xy}{N} \qquad s_X = \sqrt{\frac{\sum x^2}{N}} \qquad s_Y = \sqrt{\frac{\sum y^2}{N}} \tag{18}$$

entonces s_X y s_Y se reconocerán como las desviaciones estándar de *X* y de *Y*, respectivamente, y s_X^2 y s_Y^2 son las varianzas. La nueva cantidad s_{XY} es la *covarianza* de *X* y *Y*. En términos de la fórmula (*18*), la fórmula (*17*) puede expresarse como

$$r = \frac{s_{XY}}{s_X s_Y} \tag{19}$$

Obsérvese que *r* no sólo es independiente de las unidades de *X* y de *Y*, sino también de la elección del origen.

FÓRMULAS SIMPLIFICADAS PARA EL CÁLCULO

La fórmula (*17*) puede expresarse de la siguiente manera equivalente

$$r = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{[n \sum X^2 - (\sum X)^2][N \sum Y^2 - (\sum Y)^2]}} \tag{20}$$

con frecuencia empleada para el cálculo de *r*.

Para datos agrupados como los de una *tabla de frecuencias bivariadas* o *distribución de frecuencias bivariadas* (ver problema 14.17), conviene emplear un *método de compilación* como los de capítulos anteriores. En ese caso, la fórmula (*20*) puede expresarse

$$r = \frac{N \sum f u_X u_Y - (\sum f_X u_X)(\sum f_Y u_Y)}{\sqrt{[N \sum f_X u_X^2 - (\sum f_X u_X)^2][N \sum f_Y u_Y^2 - (\sum f_Y u_Y)^2]}} \tag{21}$$

(ver problema 14.18). Cuando se emplea esta fórmula, para facilitar los cálculos se emplea una *tabla de correlación* (ver problema 14.19).

En el caso de datos agrupados, las fórmulas (*18*) se pueden expresar como

$$s_{XY} = c_X c_Y \left[\frac{\sum f u_X u_Y}{N} - \left(\frac{\sum f_X u_X}{N} \right) \left(\frac{\sum f_Y u_Y}{N} \right) \right] \tag{22}$$

$$s_X = c_X \sqrt{\frac{\sum f_X u_X^2}{N} - \left(\frac{\sum f_X u_X}{N} \right)^2} \tag{23}$$

$$s_Y = c_Y \sqrt{\frac{\sum f_Y u_Y^2}{N} - \left(\frac{\sum f_Y u_Y}{N} \right)^2} \tag{24}$$

donde c_X y c_Y son las amplitudes de los intervalos de clase (que se suponen constantes) correspondientes a las variables *X* y *Y*, respectivamente. Obsérvese que las fórmulas (*23*) y (*24*) son equivalentes a la fórmula (*11*) del capítulo 4.

Empleando las fórmulas (*22*) y (*24*), la fórmula (*19*) parece ser equivalente a la fórmula (*21*).

RECTAS DE REGRESIÓN Y EL COEFICIENTE DE CORRELACIÓN LINEAL

La ecuación de la recta de regresión por mínimos cuadrados $Y = a_0 + a_1X$, la recta de regresión de Y sobre X, puede expresarse como

$$Y - \bar{Y} = \frac{rs_Y}{s_X}(X - \bar{X}) \qquad \text{o bien} \qquad y = \frac{rs_Y}{s_X}x \tag{25}$$

De igual manera, la recta de regresión de X sobre Y, $X = b_0 + b_1Y$, puede expresarse como

$$X - \bar{X} = \frac{rs_X}{s_Y}(Y - \bar{Y}) \qquad \text{o bien} \qquad x = \frac{rs_X}{s_Y}y \tag{26}$$

Las pendientes de las rectas de regresión (*25*) y (*26*) son iguales si y sólo si $r = \pm 1$. En esos casos las dos rectas son idénticas y existe una perfecta correlación entre X y Y. Si $r = 0$, las rectas forman ángulos rectos y no hay correlación lineal entre X y Y. Por lo tanto, el coeficiente de correlación lineal mide qué tanto se apartan las dos rectas de regresión.

Obsérvese que si las ecuaciones (*25*) y (*26*) se expresan como $Y = a_0 + a_1X$ y $X = b_0 + b_1Y$, respectivamente, entonces $a_1b_1 = r^2$ (ver problema 14.22).

CORRELACIÓN DE SERIES DE TIEMPO

Si las variables X y Y dependen del tiempo, es posible que entre X y Y exista una relación, aunque esta relación no sea, necesariamente, de dependencia directa y produzca una "correlación sin sentido". El coeficiente de correlación se obtiene considerando los pares de valores (X, Y) correspondientes a los distintos tiempos y procediendo como de costumbre, haciendo uso de las fórmulas anteriores (ver problema 14.28).

También se puede tratar de correlacionar los valores de una variable X en cierto tiempo con los correspondientes valores de X en un tiempo anterior. A esta correlación se le llama *autocorrelación*.

CORRELACIÓN DE ATRIBUTOS

Los métodos descritos en este capítulo no permiten considerar la correlación entre variables, por naturaleza, no numéricas; por ejemplo, *atributos* de individuos (como color de pelo, color de ojos, etc.). La correlación de atributos se analiza en el capítulo 12.

TEORÍA MUESTRAL DE LA CORRELACIÓN

Los N pares de valores (X, Y) de dos variables pueden considerarse como muestras de una población que consta de todos estos pares. Como hay dos variables, a esta población se le llama *población bivariada*, la que se supondrá tiene una *distribución normal bivariada*.

Se puede pensar que existe un coeficiente de correlación poblacional teórico, denotado ρ, que se estima por el coeficiente de correlación muestral r. Las pruebas de significancia o de hipótesis relacionadas con los diferentes valores de ρ requieren del conocimiento de la distribución muestral de r. Para $\rho = 0$ esta distribución es simétrica y se usa un estadístico que implica la distribución de Student. Para $\rho \neq 0$ esta distribución es sesgada; en ese caso, una transformación desarrollada por Fischer da un estadístico que está distribuido en forma aproximadamente normal. Las pruebas siguientes resumen los procedimientos empleados:

1. **Prueba de hipótesis** $\rho = 0$. Aquí se emplea el hecho de que el estadístico

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{27}$$

tiene una distribución de Student con $\nu = N - 2$ grados de libertad (ver problemas 14.31 y 14.32).

2. **Prueba de hipótesis** $\rho = \rho_0 \neq 0$. Aquí se emplea el hecho de que el estadístico

$$Z = \tfrac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) = 1.1513\log_{10}\left(\frac{1+r}{1-r}\right) \tag{28}$$

donde $e = 2.71828\ldots$, está distribuido de manera casi normal, con media y desviación estándar dadas por

$$\mu_Z = \tfrac{1}{2}\log_e\left(\frac{1+\rho_0}{1-\rho_0}\right) = 1.1513\log_{10}\left(\frac{1+\rho_0}{1-\rho_0}\right) \qquad \sigma_Z = \frac{1}{\sqrt{N-3}} \tag{29}$$

Las ecuaciones (*28*) y (*29*) también pueden usarse para hallar los límites de confianza para los coeficientes de correlación (ver problemas 14.33 y 14.34). La ecuación (*28*) se llama *transformación Z de Fischer*.

3. **Significancia de una diferencia entre coeficientes de correlación.** Para determinar si dos coeficientes de correlación r_1 y r_2, obtenidos de muestras de tamaños N_1 y N_2, respectivamente, difieren de manera notable uno de otro, empleando la ecuación (*28*) se calculan los valores Z_1 y Z_2 correspondientes a r_1 y r_2. Después se usa el hecho de que el estadístico de prueba

$$z = \frac{Z_1 - Z_2 - \mu_{Z_1 - Z_2}}{\sigma_{Z_1 - Z_2}} \tag{30}$$

donde $$\mu_{Z_1 - Z_2} = \mu_{Z_1} - \mu_{Z_2}$$

y $$\sigma_{Z_1 - Z_2} = \sqrt{\sigma_{Z_1}^2 + \sigma_{Z_2}^2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$$

está distribuido en forma normal (ver problema 14.35).

TEORÍA MUESTRAL DE LA REGRESIÓN

La ecuación de regresión $Y = a_0 + a_1X$ se obtiene basándose en datos muestrales. Se desea conocer la correspondiente ecuación de regresión para la población de la que se obtuvo la muestra. A continuación se presentan tres pruebas relacionadas con esta población:

1. **Prueba de hipótesis** $a_1 = A_1$. Para probar la hipótesis de que el coeficiente de regresión a_1 es igual a algún valor dado A_1, se emplea el hecho de que el estadístico

$$t = \frac{a_1 - A_1}{s_{Y.X}/s_X}\sqrt{N-2} \tag{31}$$

tiene una distribución de Student con $N - 2$ grados de libertad. Esto también se puede emplear para hallar intervalos de confianza para los coeficientes de regresión poblacional a partir de valores muestrales (ver los problemas 14.36 y 14.37).

2. **Prueba de la hipótesis para valores pronosticados.** Sea Y_0 el valor pronosticado para Y, correspondiente a $X = X_0$, mediante la ecuación de regresión muestral (es decir, $Y_0 = a_0 + a_1X_0$). Sea Y_P el valor pronosticado para Y que corresponde a $X = X_0$ en la población. Entonces, el estadístico

$$t = \frac{Y_0 - Y_p}{s_{Y.X}\sqrt{N + 1 + (X_0 - \bar{X})^2/s_X^2}}\sqrt{N-2} = \frac{Y_0 - Y_p}{\hat{s}_{X.Y}\sqrt{1 + 1/N + (X_0 - \bar{X})^2/(Ns_X^2)}} \tag{32}$$

tiene una distribución de Student con $N - 2$ grados de libertad. A partir de esta fórmula se pueden hallar límites de confianza para valores poblacionales pronosticados (ver el problema 14.38).

3. **Prueba de hipótesis para valores pronosticados para la media.** Sea Y_0 el valor pronosticado para Y, correspondiente a $X = X_0$, empleando la ecuación de regresión muestral (es decir, $Y_0 = a_0 + a_1X_0$). Sea $\bar{Y}_p$ el *valor medio* pronosticado de Y que corresponde a $X = X_0$ en la población. Entonces el estadístico

$$t = \frac{Y_0 - \bar{Y}_p}{s_{Y.X}\sqrt{1 + (X_0 - \bar{X})^2/s_X^2}}\sqrt{N-2} = \frac{Y_0 - \bar{Y}_p}{\hat{s}_{Y.X}\sqrt{1/N + (X_0 - \bar{X})^2/(Ns_X^2)}} \tag{33}$$

tiene una distribución de Student con $N - 2$ grados de libertad. A partir de esta fórmula se pueden hallar límites de confianza para valores pronosticados para la media poblacional (ver el problema 14.39).

PROBLEMAS RESUELTOS

DIAGRAMAS DE DISPERSIÓN Y RECTAS DE REGRESIÓN

14.1 En la tabla 14.1 X y Y son las estaturas de 12 padres y de sus hijos mayores.

a) Con estos datos, construir un diagrama de dispersión.
b) Resolviendo las ecuaciones normales, encontrar la línea de regresión de mínimos cuadrados correspondiente a la estatura del padre sobre la estatura del hijo. También encontrar esta línea empleando SPSS.
c) Resolviendo las ecuaciones normales, encontrar la línea de regresión de mínimos cuadrados correspondiente a la estatura del hijo sobre la estatura del padre. Encontrar también esta línea empleando STATISTIX.

Tabla 14.1

Estatura X del padre (in)	65	63	67	64	68	62	70	66	68	67	69	71
Estatura Y del hijo (in)	68	66	68	65	69	66	68	65	71	67	68	70

SOLUCIÓN

a) El diagrama de dispersión se obtiene graficando los puntos (X, Y) en un sistema de coordenadas rectangulares, como el que se muestra en la figura 14-2.

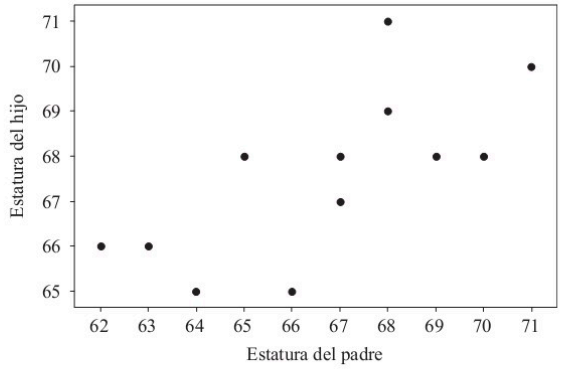


Figura 14-2 Diagrama de dispersión de los datos de la tabla 14.1.

b) La recta de regresión de Y sobre X es $Y = a_0 + a_1X$, donde a_0 y a_1 se obtienen resolviendo las ecuaciones normales.

$$\sum Y = a_0N + a_1 \sum X$$
$$\sum XY = a_0 \sum X + a_1 \sum X^2$$

En la tabla 14.2 se presentan las sumas a partir de las cuales las ecuaciones normales son

$$12a_0 + 800a_1 = 811$$
$$800a_0 + 53\,418a_1 = 54\,107$$

de donde se encuentra que $a_0 = 35.82$ y $a_1 = 0.476$, con lo que $Y = 35.82 + 0.476X$.

A continuación se presenta parte del resultado que se obtiene con la secuencia **Analyze** → **Regresión** → **Linear** de SPSS.

Coeficientes[a]

Modelo		Coeficientes sin estandarizar		Coeficientes estandarizados		
		B	Error estándar	Beta	t	Sig
1	(Constante)	35.825	10.178		3.520	.006
	Estpadre	.476	.153	.703	3.123	.011

[a]Variable dependiente: Esthijo.

Delante de la palabra (`Constante`) se encuentra el valor de a_0 y delante de la palabra `Estpadre` se encuentra el valor de a_1.

Tabla 14.2

X	Y	X^2	XY	Y^2
65	68	4 225	4 420	4 624
63	66	3 969	4 158	4 356
67	68	4 489	4 556	4 624
64	65	4 096	4 160	4 225
68	69	4 624	4 692	4 761
62	66	3 844	4 092	4 356
70	68	4 900	4 760	4 624
66	65	4 356	4 290	4 225
68	71	4 624	4 828	5 041
67	67	4 489	4 489	4 489
69	68	4 761	4 692	4 624
71	70	5 041	4 970	4 900
$\sum X = 800$	$\sum Y = 811$	$\sum X^2 = 53\,418$	$\sum XY = 54\,107$	$\sum Y^2 = 54\,849$

c) La recta de regresión de X sobre Y es $X = b_0 + b_1 Y$, donde b_0 y b_1 se obtienen resolviendo las ecuaciones normales

$$\sum X = b_0 N + b_1 \sum Y$$
$$\sum XY = b_0 \sum Y + b_1 \sum Y^2$$

Empleando las sumas de la tabla 14.2 estas ecuaciones son:

$$12b_0 + 811b_1 = 800$$
$$811b_0 + 54\,849b_1 = 54\,107$$

de las cuales se encuentra que $b_0 = -3.38$ y $b_1 = 1.036$, por lo que $X = -3.38 + 1.036Y$

A continuación se presenta parte del resultado que se obtiene con la secuencia **Statistics** → **Linear models** → **Linear regresión** de STATISTIX:

```
Statistix 8.0
Unweighted Least Squares Linear Regression of Htfather

Predictor
Variable        Coefficient     Std Error        T          P
Constant         -3.37687        22.4377       -0.15      0.8834
Htson             1.03640         0.33188       3.12      0.0108
```

Delante de la palabra `constant` se encuentra el valor $b_0 = -3.37687$ y delante de la palabra `Esthijo` se encuentra el valor $b_1 = 1.0364$.

14.2 Resolver el problema 14.1 usando MINITAB. Construir tablas en las que se den los valores ajustados, Y_{est}, y los residuales. Encontrar la suma de los cuadrados de los residuales correspondientes a estas dos rectas de regresión.

SOLUCIÓN

Primero se hallará la línea de regresión por mínimos cuadrados de Y sobre X. A continuación se muestran parte de los resultados que da MINITAB. En la tabla 14.3 se dan los valores ajustados, los residuales y los cuadrados de los residuales correspondientes a la línea de regresión de Y sobre X.

Tabla 14.3

X	Y	Valor ajustado Y_{est}	Residual $Y - Y_{est}$	Cuadrado del residual
65	68	66.79	1.21	1.47
63	66	65.84	0.16	0.03
67	68	67.74	0.26	0.07
64	65	66.31	–1.31	1.72
68	69	68.22	0.78	0.61
62	66	65.36	0.64	0.41
70	68	69.17	–1.17	1.37
66	65	67.27	–2.27	5.13
68	71	68.22	2.78	7.74
67	67	67.74	–0.74	0.55
69	68	68.69	–0.69	0.48
71	70	69.65	0.35	0.12
			Suma = 0	Suma = 19.70

```
MTB > Regress 'Y' on 1 predictor 'X'
```

Análisis de regresión

```
La ecuación de regresión es Y = 35.8 + 0.476 X
```

El resultado que da MINITAB al hallar la línea de regresión por mínimos cuadrados de X sobre Y es el siguiente:

```
MTB > Regress 'X' on 1 predictor 'Y'
```

Análisis de regresión

```
La ecuación de regresión es X = -3.4 + 1.04 Y
```

En la tabla 14.4 se dan los valores ajustados, los residuales y los cuadrados de los residuales correspondientes a la línea de regresión de X sobre Y.

Tabla 14.4

X	Y	Valor ajustado X_{est}	Residual $X - X_{est}$	Cuadrado del residual
65	68	67.10	−2.10	4.40
63	66	65.03	−2.03	4.10
67	68	67.10	−0.10	0.01
64	65	63.99	0.01	0.00
68	69	68.13	−0.13	0.02
62	66	65.03	−3.03	9.15
70	68	67.10	2.90	8.42
66	65	63.99	2.01	4.04
68	71	70.21	−2.21	4.87
67	67	66.06	0.94	0.88
69	68	67.10	1.90	3.62
71	70	69.17	1.83	3.34
			Suma = 0	Suma = 42.85

Comparando la suma de cuadrados de los residuales se ve que el ajuste de la recta de regresión de mínimos cuadrados de Y sobre X es mucho mejor que el ajuste de la recta de regresión de mínimos cuadrados de X sobre Y. Recuérdese que cuanto menor sea la suma de los cuadrados de los residuales, el modelo de regresión se ajusta mejor a los datos. La estatura del padre es mejor predictor de la estatura del hijo que la estatura del hijo de la estatura del padre.

ERROR ESTÁNDAR DE ESTIMACIÓN

14.3 Si la línea de regresión de Y sobre X está dada por $Y = a_0 + a_1X$, probar que el error estándar de estimación $s_{Y.X}$ está dado por

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N}$$

SOLUCIÓN

Los valores estimados para Y, de acuerdo con la línea de regresión, están dados por $Y_{est} = a_0 + a_1X$. Por lo tanto,

$$s_{Y.X}^2 = \frac{\sum (Y - Y_{est})^2}{N} = \frac{\sum (Y - a_0 - a_1X)^2}{N}$$
$$= \frac{\sum Y(Y - a_0 - a_1X) - a_0 \sum (Y - a_0 - a_1X) - a_1 \sum X(Y - a_0 - a_1X)}{N}$$

Pero
$$\sum (Y - a_0 - a_1X) = \sum Y - a_0N - a_1 \sum X = 0$$

y
$$\sum X(Y - a_0 - a_1X) = \sum XY - a_0 \sum X - a_1 \sum X^2 = 0$$

ya que de acuerdo con las ecuaciones normales

$$\sum Y = a_0N + a_1 \sum X$$
$$\sum XY = a_0 \sum X + a_1 \sum X^2$$

Por lo tanto,
$$s_{Y.X}^2 = \frac{\sum Y(Y - a_0 - a_1X)}{N} = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N}$$

Este resultado puede extenderse a ecuaciones de regresión no lineales.

14.4 Si $x = X - \bar{X}$ y $y = Y - \bar{Y}$, mostrar que la ecuación del problema 14.3 puede expresarse

$$s_{Y.X}^2 = \frac{\sum y^2 - a_1 \sum xy}{N}$$

SOLUCIÓN

De acuerdo con el problema 14.3, si $X = x + \bar{X}$ y $Y = y + \bar{Y}$, se tiene

$$\begin{aligned} Ns_{Y.X}^2 &= \sum Y^2 - a_0 \sum Y - a_1 \sum XY = \sum (y + \bar{Y})^2 - a_0 \sum (y + \bar{Y}) - a_1 \sum (x + \bar{X})(y + \bar{Y}) \\ &= \sum (y^2 + 2y\bar{Y} + \bar{Y}^2) - a_0(\sum y + N\bar{Y}) - a_1 \sum (xy + \bar{X}y + x\bar{Y} + \bar{X}\bar{Y}) \\ &= \sum y^2 + 2\bar{Y} \sum y + N\bar{Y}^2 - a_0 N\bar{Y} - a_1 \sum xy - a_1\bar{X} \sum y - a_1\bar{Y} \sum x - a_1 N\bar{X}\bar{Y} \\ &= \sum y^2 + N\bar{Y}^2 - a_0 N\bar{Y} - a_1 \sum xy - a_1 N\bar{X}\bar{Y} \\ &= \sum y^2 - a_1 \sum xy + N\bar{Y}(\bar{Y} - a_0 - a_1\bar{X}) \\ &= \sum y^2 - a_1 \sum xy \end{aligned}$$

donde se han empleado los resultados $\sum x = 0$, $\sum y = 0$ y $\bar{Y} = a_0 + a_1\bar{X}$ (que se obtienen al dividir entre N ambos lados de la ecuación normal $\sum Y = a_0 N + a_1 \sum X$ por N).

14.5 Dados los datos del problema 14.1, calcular el error estándar de estimación $s_{Y.X}$ empleando: *a*) la definición y *b*) la ecuación obtenida en el problema 14.4.

SOLUCIÓN

a) De acuerdo con el problema 14.1*b*), la recta de regresión de Y sobre X es $Y = 35.82 + 0.476X$. En la tabla 14.5 se dan los valores reales de Y (tomados de la tabla 14.1) y los valores estimados de Y, que se denotan Y_{est}, obtenidos empleando la recta de regresión; por ejemplo, para $X = 65$ se tiene $Y_{\text{est}} = 35.82 + 0.476(65) = 66.76$. También se dan los valores $Y - Y_{\text{est}}$, que se necesitan para calcular $s_{Y.X}$:

$$s_{Y.X}^2 = \frac{\sum (Y - Y_{\text{est}})}{N} = \frac{(1.24)^2 + (0.19)^2 + \cdots + (0.38)^2}{12} = 1.642$$

y $s_{Y.X} = \sqrt{1.1642} = 1.28$ in.

b) De acuerdo con los problemas 14.1, 14.2 y 14.4

$$s_{Y.X}^2 = \frac{\sum y^2 - a_1 \sum xy}{N} = \frac{38.92 - 0.476(40.34)}{12} = 1.643$$

y $s_{Y.X} = \sqrt{1.643} = 1.28$ in.

Tabla 14.5

X	65	63	67	64	68	62	70	66	68	67	69	71
Y	68	66	68	65	69	66	68	65	71	67	68	70
Y_{est}	66.76	65.81	67.71	66.28	68.19	65.33	69.14	67.24	68.19	67.71	68.66	69.62
$Y - Y_{\text{est}}$	1.24	0.19	0.29	−1.28	0.81	0.67	−1.14	−2.24	2.81	−0.71	−0.66	0.38

14.6 *a*) Construir dos rectas que sean paralelas a la recta de regresión del problema 14.1 y que se encuentren a una distancia vertical $s_{Y.X}$ de ella.

b) Determinar el porcentaje de los datos que caen entre estas dos líneas.

SOLUCIÓN

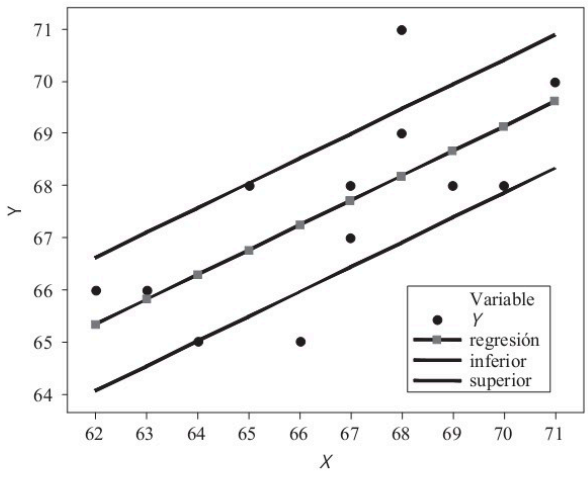


Figura 14-3 De los datos, el 66% se encuentra a una distancia no mayor a $S_{Y.X}$ de la línea de regresión.

a) La recta de regresión $Y = 35.82 + 0.476X$, obtenida en el problema 14.1, es la recta que aparece marcada con los rombos. Es la recta de enmedio de las tres rectas que aparecen en la figura 14-3; hay otras dos rectas que se encuentran cada una a una distancia $S_{Y.X} = 1.28$ de la recta de regresión. A estas rectas se les llama rectas inferior y superior.

b) En la figura 14-3, los datos aparecen como círculos en negro. Ocho de los 12 datos, es decir el 66.7%, se encuentran entre las rectas inferior y superior. Dos datos se encuentran fuera de estas rectas y otros dos se hallan sobre estas rectas.

VARIACIÓN EXPLICADA Y VARIACIÓN NO EXPLICADA

14.7 Probar que $\sum (Y - \bar{Y})^2 = \sum (Y - Y_{\text{est}})^2 + \sum (Y_{\text{est}} - \bar{Y})^2$.

SOLUCIÓN

Elevando al cuadrado ambos lados de $Y - \bar{Y} = (Y - Y_{\text{est}}) + (Y_{\text{est}} - \bar{Y})$ y sumando después, se tiene

$$\sum (Y - \bar{Y})^2 = \sum (Y - Y_{\text{est}})^2 + \sum (Y_{\text{est}} - \bar{Y})^2 + 2 \sum (Y - Y_{\text{est}})(Y_{\text{est}} - \bar{Y})$$

La ecuación buscada se obtiene inmediatamente si se demuestra que la última suma es cero; en el caso de la regresión lineal, esto es así debido a que

$$\begin{aligned}\sum (Y - Y_{\text{est}})(Y_{\text{est}} - \bar{Y}) &= \sum (Y - a_0 - a_1 X)(a_0 + a_1 X - \bar{Y}) \\ &= a_0 \sum (Y - a_0 - a_1 X) + a_1 \sum X(Y - a_0 - a_1 X) - \bar{Y} \sum (Y - a_0 - a_1 X) = 0\end{aligned}$$

y por las ecuaciones normales, $\sum (Y - a_0 - a_1 X) = 0$ y $\sum X(Y - a_0 - a_1 X) = 0$.

De igual manera, empleando la curva de mínimos cuadrados dada por $Y_{\text{est}} = a_0 + a_1 X + a_2 X^2 + \cdots + a_n X^n$, puede mostrarse que este resultado también es válido para la regresión no lineal.

14.8 Dados los datos del problema 14.1, calcular: *a*) la variación total, *b*) la variación no explicada y *c*) la variación explicada.

SOLUCIÓN

La recta de regresión por mínimos cuadrados es $Y_{\text{est}} = 35.8 + 0.476X$. En la tabla 14.6 se ve que la variación total $= \sum (Y - \bar{Y})^2 = 38.917$, la variación no explicada $= \sum (Y - Y_{\text{est}})^2 = 19.703$ y la variación explicada $= \sum (Y_{\text{est}} - \bar{Y})^2 = 19.214$.

Tabla 14.6

Y	Y_{est}	$(Y-\bar{Y})^2$	$(Y-Y_{est})^2$	$(Y_{est}-\bar{Y})^2$
68	66.7894	0.1739	1.46562	0.62985
66	65.8366	2.5059	0.02669	3.04986
68	67.7421	0.1739	0.06650	0.02532
65	66.3130	6.6719	1.72395	1.61292
69	68.2185	2.0079	0.61074	0.40387
66	65.3602	2.5059	0.40930	4.94068
68	69.1713	0.1739	1.37185	2.52257
65	67.2657	6.6719	5.13361	0.10065
71	68.2185	11.6759	7.73672	0.40387
67	67.7421	0.3399	0.55075	0.02532
68	68.6949	0.1739	0.48286	1.23628
70	69.6476	5.8419	0.12416	4.26273
$\bar{Y} = 67.5833$		Suma = 38.917	Suma = 19.703	Suma = 19.214

Los siguientes resultados de MINITAB dan las mismas sumas de cuadrados. Estas sumas aparecen en negritas. Obsérvese la enorme cantidad de cálculos que este software le ahorra al usuario.

```
MTB > Regress 'Y' 1 'X';
SUBC> Constant;
SUBC> Brief 1.
```

Análisis de regresión

```
The regression equation is
Y = 35.8 + 0.476 X

Analysis of Variance

Source            DF          SS          MS        F        P
Regression         1      19.214      19.214     9.75    0.011
Residual Error    10      19.703       1.970
Total             11      38.917
```

COEFICIENTE DE CORRELACIÓN

14.9 Usar los resultados del problema 14.8 para hallar: *a*) el coeficiente de determinación y *b*) el coeficiente de correlación.

SOLUCIÓN

a) Coeficiente de determinación $= r^2 = \dfrac{\text{variación explicada}}{\text{variación total}} = \dfrac{19.214}{38.917} = 0.4937$

b) Coeficiente de correlación $= r = \pm\sqrt{0.4937} = \pm 0.7027$

Como X y Y se relacionan en forma directa, se elige el signo positivo. A dos lugares decimales $r = 0.70$.

14.10 Probar que para la regresión lineal, el coeficiente de correlación entre las variables X y Y puede expresarse como

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

donde $x = X - \bar{X}$ y $y = Y - \bar{Y}$.

SOLUCIÓN

La recta de regresión por mínimos cuadrados de Y sobre X puede expresarse $Y_{est} = a_0 + a_1X$ o bien $y_{est} = a_1x$, donde [ver problema 13.15*a*)]

$$a_1 = \frac{\sum xy}{\sum x^2} \qquad \text{y} \qquad y_{est} = Y_{est} - \bar{Y}$$

Entonces

$$r^2 = \frac{\text{variación explicada}}{\text{variación total}} = \frac{\sum (Y_{est} - \bar{Y})^2}{\sum (Y - \bar{Y})^2} = \frac{\sum y_{est}^2}{\sum y^2}$$

$$= \frac{\sum a_1^2 x^2}{\sum y^2} = \frac{a_1^2 \sum x^2}{\sum y^2} = \left(\frac{\sum xy}{\sum x^2}\right)^2 \frac{\sum x^2}{\sum y^2} = \frac{(\sum xy)^2}{(\sum x^2)(\sum y^2)}$$

y

$$r = \pm \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

Sin embargo, como la cantidad

$$\frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

es positiva cuando y_{est} aumenta a medida que x aumenta (es decir, correlación lineal positiva) y negativa cuando y_{est} disminuye a medida que x aumenta (es decir, correlación lineal negativa), esta expresión tiene *automáticamente* el signo correcto. Por lo tanto, el coeficiente de correlación lineal se define como

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

A esta expresión se le conoce como *fórmula producto-momento* para el coeficiente de correlación lineal.

FÓRMULA PRODUCTO-MOMENTO PARA EL COEFICIENTE DE CORRELACIÓN LINEAL

14.11 Encontrar el coeficiente de correlación lineal entre las variables X y Y que se presentan en la tabla 14.7.

Tabla 14.7

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

SOLUCIÓN

Para facilitar los cálculos se elabora la tabla 14.8.

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{84}{\sqrt{(132)(56)}} = 0.977$$

Esto indica que existe una correlación lineal muy elevada entre estas variables, como ya se observó en los problemas 13.8 y 13.12.

Tabla 14.8

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy	y^2
1	1	−6	−4	36	24	16
3	2	−4	−3	16	12	9
4	4	−3	−1	9	3	1
6	4	−1	−1	1	1	1
8	5	1	0	1	0	0
9	7	2	2	4	4	4
11	8	4	3	16	12	9
14	9	7	4	49	28	16
$\sum X = 56$ $\bar{X} = 56/8 = 7$	$\sum Y = 40$ $\bar{Y} = 40/8 = 5$			$\sum x^2 = 132$	$\sum xy = 84$	$\sum y^2 = 56$

14.12 Con objeto de investigar la relación entre el promedio de calificaciones y la cantidad de horas por semana que se ve televisión, se recolectan los datos que se muestran en la tabla 14.9 y en la figura 14-4, y se emplea EXCEL para obtener un diagrama de dispersión de los datos. La información corresponde a 10 estudiantes de secundaria, X es la cantidad de horas por semana que el estudiante ve televisión (horas de TV) y Y es su promedio de calificaciones.

Tabla 14.9

Horas de TV	Promedio de calificaciones
20	2.35
5	3.8
8	3.5
10	2.75
13	3.25
7	3.4
13	2.9
5	3.5
25	2.25
14	2.75

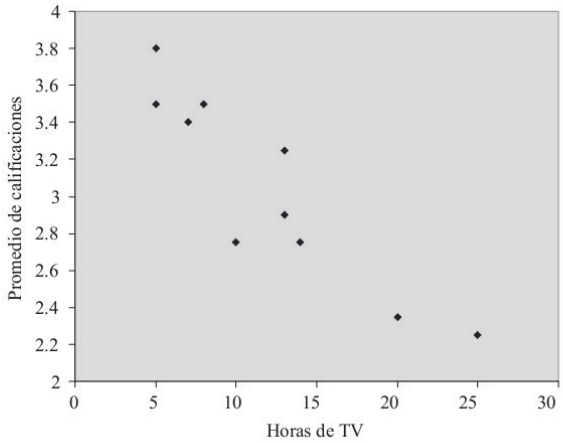


Figura 14-4 EXCEL, diagrama de dispersión de datos del problema 14.12.

Usar EXCEL para calcular el coeficiente de correlación de estas dos variables y verificar empleando la fórmula de producto-momento.

SOLUCIÓN

Para hallar el coeficiente de correlación se emplea la función de EXCEL `=CORREL(E2:E11,F2:F11)`, ingresando en las celdas `E2:E11` las horas de televisión y en las celdas `F2:F11` los promedios de calificaciones. El coeficiente de correlación es -0.9097. El signo negativo indica que las dos variables están inversamente correlacionadas. Es decir, cuanto mayor es la cantidad de horas que se ve televisión, menor es el promedio de calificaciones.

14.13 En un estudio se registran los salarios iniciales (en miles), Y, y los años de estudio, X, de 10 empleados. En la tabla 14.10 y en la figura 14-5 se presentan los datos y una gráfica de dispersión empleando SPSS.

Tabla 14.10

Salario inicial	Años de estudio
35	12
46	16
48	16
50	15
40	13
65	19
28	10
37	12
49	17
55	14

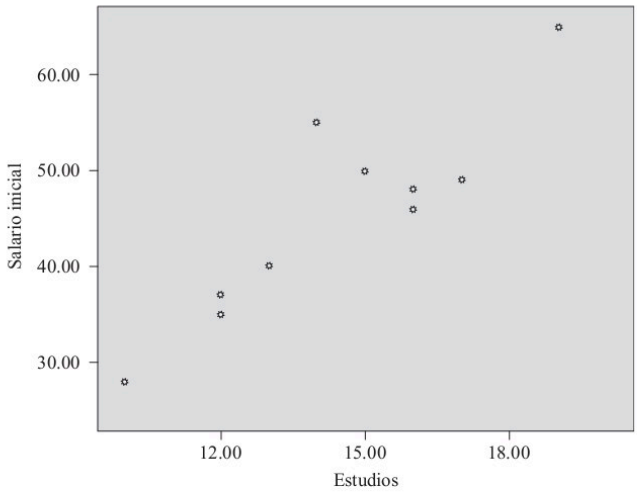


Figura 14-5 SPSS, diagrama de dispersión del problema 14.13.

Usar SPSS para calcular el coeficiente de correlación de estas dos variables y verificar usando la fórmula del producto-momento.

SOLUCIÓN

Correlaciones

		salario inicial	estudios
salario inicial	Correlación de Pearson	1	.891**
	Sig. (2 colas)		.001
	N	10	10
estudios	Correlación de Pearson	.891**	1
	Sig. (2 colas)	.001	
	N	10	10

**Correlación significativa al nivel 0.001 (2 colas).

La secuencia **Analyze** → **Correlate** → **Bivariate** de SPSS da la correlación empleando la fórmula del producto-momento. A esta fórmula también se le llama correlación de Pearson.

El resultado da el coeficiente de correlación $r = 0.891$.

14.14 En un estudio realizado con 10 estudiantes se registró la cantidad de horas por semana que emplean su teléfono celular, Y, y la cantidad de letras en su nombre, X. En la tabla 14.11 y en la figura 14-6 se presentan los datos y el diagrama de dispersión obtenido con STATISTIX.

Tabla 14.11

Horas de celular	Letras en el nombre
6	13
6	11
3	12
17	7
19	14
14	4
15	4
3	13
13	4
7	9

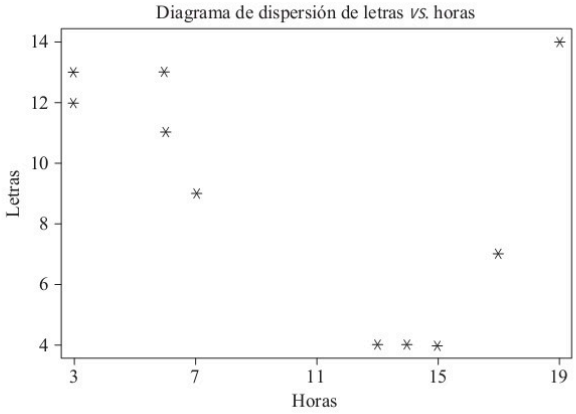


Figura 14-6 STATISTIX, diagrama de dispersión de los datos de la tabla 14.11.

Usar STATISTIX para calcular el coeficiente de correlación de las dos variables y verificar usando la fórmula del producto-momento.

SOLUCIÓN

Con la secuencia "**Statistics** → **Linear models** → **correlations (Pearson)**" se obtiene el resultado siguiente:

```
Statistix 8.0
Correlations (Pearson)

                    Hours
Letters           -0.4701
P-VALUE            0.1704
```

El coeficiente de correlación es $r = -0.4701$. Entre estas dos variables no existe una correlación significativa.

14.15 Mostrar que el coeficiente de correlación lineal está dado por

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

SOLUCIÓN

Si se escribe $x = X - \bar{X}$ y $y = Y - \bar{Y}$ en la fórmula del problema 14.10, se tiene

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{[\sum (X - \bar{X})^2][\sum (Y - \bar{Y})^2]}} \tag{34}$$

Pero
$$\begin{aligned}\sum (X - \bar{X})(Y - \bar{Y}) &= \sum (XY - \bar{X}Y - X\bar{Y} + \bar{X}\bar{Y}) = \sum XY - \bar{X}\sum Y - \bar{Y}\sum X + N\bar{X}\bar{Y} \\ &= \sum XY - N\bar{X}\bar{Y} - N\bar{Y}\bar{X} + N\bar{X}\bar{Y} = \sum XY - N\bar{X}\bar{Y} \\ &= \sum XY - \frac{(\sum X)(\sum Y)}{N}\end{aligned}$$

ya que $\bar{X} = (\sum X)/N$ y $\bar{Y} = (\sum Y)/N$. De igual manera,

$$\begin{aligned}\sum (X - \bar{X})^2 &= \sum (X^2 - 2X\bar{X} + \bar{X}^2) = \sum X^2 - 2\bar{X}\sum X + N\bar{X}^2 \\ &= \sum X^2 - \frac{2(\sum X)^2}{N} + \frac{(\sum X)^2}{N} = \sum X^2 - \frac{(\sum X)^2}{N}\end{aligned}$$

y
$$\sum (Y - \bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{N}$$

Por lo tanto, la ecuación (*34*) se convierte en

$$r = \frac{\sum XY - (\sum X)(\sum Y)/N}{\sqrt{[\sum X^2 - (\sum X)^2/N][\sum Y^2 - (\sum Y)^2/N]}} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

14.16 Se estudió la relación entre el exceso de peso y la presión sanguínea alta en adultos obesos. En la tabla 14.12 se presentan exceso de peso, en libras, y unidades superiores a 80 en la presión diastólica. En la figura 14-7 se presenta el diagrama de dispersión obtenido con SAS.

Tabla 14.12

Exceso de peso en libras	Unidades superiores a 80
75	15
86	13
88	10
125	27
75	20
30	5
47	8
150	31
114	78
68	22

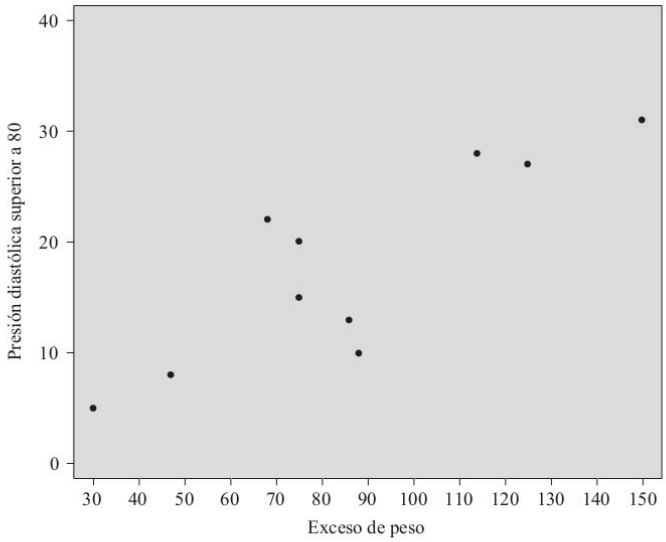


Figura 14-7 SAS, diagrama de dispersión para el problema 14.16.

Usar SAS para calcular el coeficiente de correlación de estas dos variables y verificar usando la fórmula del producto-momento.

SOLUCIÓN

Con la secuencia **Statistics** → **Descriptive** → **Correlations** de SAS se obtiene el procedimiento para la correlación, una parte del cual se muestra a continuación.

```
                The CORR Procedure
                     Overwt                 Over80
Overwt               1.00000                0.85536
Overwt                                      0.0016
Over80               0.85536                1.00000
Over80               0.0016
```

El coeficiente de correlación dado en estos resultados es 0.85536. Existe una correlación significativa entre el exceso de peso de una persona y una presión diastólica superior a 80.

COEFICIENTE DE CORRELACIÓN PARA DATOS AGRUPADOS

14.17 En la tabla 14.13 se muestran las distribuciones de frecuencias de las calificaciones finales en matemáticas y en física de 100 estudiantes. De acuerdo con esta tabla determinar.

a) El número de estudiantes que en matemáticas obtuvo una calificación entre 70 y 79, y en física una calificación entre 80 y 89.

b) El porcentaje de estudiantes cuya calificación en matemáticas es menor a 70.

c) El número de estudiantes que tiene 70 o más en física y menos de 80 en matemáticas.

d) El porcentaje de estudiantes que aprueba por lo menos una de estas dos materias, suponiendo que la calificación para aprobar es 60.

SOLUCIÓN

a) En la tabla 14.13 se desciende por la columna cuyo encabezado es 70-79 (calificación en matemáticas) hasta el renglón marcado 80-89 (calificación en física), donde la entrada es 4, que es el número de estudiantes buscado.

Tabla 14.13

		Calificación en matemáticas						
		40-49	50-59	60-69	70-79	80-89	90-99	Total
	90-99				2	4	4	10
Calificación en física	80-89			1	4	6	5	16
	70-79			5	10	8	1	24
	60-69	1	4	9	5	2		21
	50-59	3	6	6	2			17
	40-49	3	5	4				12
	Total	7	15	25	23	20	10	100

b) El número de estudiantes cuya calificación en matemáticas es menos de 70 es el número de estudiantes cuya calificación corresponde a 40-49 + el número de estudiantes cuya calificación está en 50-59 + el número de estudiantes cuya calificación se halla en 60-69 $= 7 + 15 + 25 = 47$. Por lo tanto, el porcentaje buscado es $47/100 = 47\%$.

c) El número buscado de estudiantes es la suma de las entradas en la tabla 14.14 (que presenta parte de las entradas de la tabla 14.13). Por lo tanto, el número buscado de estudiantes es $1 + 5 + 2 + 4 + 10 = 22$.

d) En la tabla 14.15 (tomada de la tabla 14.13) se muestra el número de alumnos que tiene una calificación menor a 60 en física o en matemáticas o en ambas materias, que es $3 + 3 + 6 + 5 = 17$. Por lo tanto, el número de estudiantes con una calificación de 60 o más en física o en matemáticas, o en ambas, es $100 - 17 = 83$. El porcentaje buscado es $83/100 = 83\%$.

Tabla 14.14

		Calificaciones en matemáticas	
		60-69	70-79
Calificaciones en física	90-99		2
	80-89	1	4
	70-79	5	10

Tabla 14.15

		Calificaciones en matemáticas	
		40-49	50-59
Calificaciones en física	50-59	3	6
	40-49	3	5

La tabla 14.13 a veces se denomina *tabla de frecuencias bivariada* o *distribución de frecuencias bivariada.* Cada cuadro de la tabla se llama *celda* y corresponde a un par de clases o intervalos de clase. El número indicado en la celda se conoce como *frecuencia de celda.* Por ejemplo, en la parte *a*) el número 4 es la frecuencia de la celda que corresponde al par de intervalos de clase 70-79 en matemáticas y 80-89 en física.

Los totales indicados en la última fila y la última columna se denominan *totales marginales* o *frecuencias marginales.* Corresponden, respectivamente, a las frecuencias de clase de las distribuciones de frecuencias separadas de las calificaciones de matemáticas y de física.

14.18 Mostrar cómo modificar la fórmula del problema 14.15 en el caso de datos agrupados, como en la tabla de frecuencias bivariada (tabla 14.13).

SOLUCIÓN

En el caso de datos agrupados se puede considerar que los valores de las variables X y Y coinciden con las marcas de clase y que f_X y f_Y son las correspondientes frecuencias de clase, o frecuencias marginales, que se muestran en el último renglón y en la última columna de la tabla de frecuencias bivariada. Si f representa las diversas frecuencias de celda que corresponden a los pares de marcas de clase (X, Y), entonces la fórmula del problema 14.15 puede reemplazarse por la fórmula

$$r = \frac{N\sum fXY - (\sum f_X X)(\sum f_Y Y)}{\sqrt{[N\sum f_X X^2 - (\sum f_X X)^2][N\sum f_Y Y^2 - (\sum f_Y Y)^2]}} \tag{35}$$

Si $X = A + c_x u_x$ y $Y = B + c_Y u_Y$, donde c_X y c_Y son las amplitudes de los intervalos de clase (que se suponen constantes) y A y B son dos marcas de clase cualesquiera que corresponden a estas variables, la fórmula (*35*) se convierte en la fórmula (*21*) de este capítulo:

$$r = \frac{N\sum fu_X u_Y - (\sum f_X u_X)(\sum f_Y u_Y)}{\sqrt{[N\sum f_X u_X^2 - (\sum f_X u_X)^2][N\sum f_Y u_Y^2 - (\sum f_Y u_Y)^2]}} \tag{21}$$

Éste es el *método de codificación* empleado en capítulos anteriores como método abreviado para el cálculo de medias, desviaciones estándar y momentos superiores.

14.19 Encontrar el coeficiente de correlación lineal correspondiente a las calificaciones de matemáticas y de física del problema 14.17.

SOLUCIÓN

Se usará la fórmula (*21*). Para facilitar los cálculos se elabora la tabla 14.16, a la que se le llama *tabla de correlación*. Las sumas $\sum f_X$, $\sum f_X u_X$, $\sum f_X u_X^2$, $\sum f_Y$, $\sum f_Y u_Y$ y $\sum f_Y u_Y^2$ se obtienen empleando el método de codificación, como en capítulos anteriores.

En la tabla 14.16, el número que aparece en la esquina de cada celda representa el producto $fu_X u_Y$, donde f es la frecuencia de celda. La suma de estos números de las esquinas en cada renglón se indica en el renglón correspondiente de la última columna. La suma de estos números de las esquinas en cada columna se indica en la columna correspondiente del último renglón. Los totales finales del último renglón y de la última columna son iguales y representan $\sum fu_X u_Y$.

De acuerdo con la tabla 14.16, se tiene

$$r = \frac{N\sum fu_X u_Y - (\sum f_X u_X)(\sum f_Y u_Y)}{\sqrt{[N\sum f_X u_X^2 - (\sum f_X u_X)^2][N\sum f_Y u_Y^2 - (\sum f_Y u_Y)^2]}}$$

$$= \frac{(100)(125) - (64)(-55)}{\sqrt{[(100)(236) - (64)^2][(100)(253) - (-55)^2]}} = \frac{16\,020}{\sqrt{(19\,504)(22\,275)}} = 0.7686$$

14.20 Empleando la tabla 14.16 calcular: *a*) s_X, *b*) s_Y y *c*) s_{XY} y verificar la fórmula $r = s_{XY}/(s_X s_Y)$.

SOLUCIÓN

a) $$s_X = c_X\sqrt{\frac{\sum f_X u_X^2}{N} - \left(\frac{\sum f_X u_X}{N}\right)^2} = 10\sqrt{\frac{236}{100} - \left(\frac{64}{100}\right)^2} = 13.966$$

b) $$s_Y = c_Y\sqrt{\frac{\sum f_Y u_Y^2}{N} - \left(\frac{\sum f_Y u_Y}{N}\right)^2} = 10\sqrt{\frac{253}{100} - \left(\frac{-55}{100}\right)^2} = 14.925$$

c) $$s_{XY} = c_X c_Y\left[\frac{\sum fu_X u_Y}{N} - \left(\frac{\sum f_X u_X}{N}\right)\left(\frac{\sum f_Y u_Y}{N}\right)\right] = (10)(10)\left[\frac{125}{100} - \left(\frac{64}{100}\right)\left(\frac{-55}{100}\right)\right] = 160.20$$

Por lo tanto, la desviación estándar de las calificaciones de matemáticas y de física son 14.0 y 14.9, respectivamente, y la covarianza es 160.2. Por lo tanto, el coeficiente de correlación r es

Tabla 14.16

Calificaciones en física Y		Calificaciones en matemáticas X										
	X	44.5	54.5	64.5	74.5	84.5	94.5				
Y	u_X / u_Y	-2	-1	0	1	2	3	f_Y	$f_Y u_Y$	$f_Y u_Y^2$	Suma de los números en las esquinas de cada renglón
94.5	2				2 [4]	4 [16]	4 [24]	10	20	40	44
84.5	1			1 [0]	4 [4]	6 [12]	5 [15]	16	16	16	31
74.5	0			5 [0]	10 [0]	8 [0]	1 [0]	24	0	0	0
64.5	-1	1 [2]	4 [4]	9 [0]	5 [-5]	2 [-4]		21	-21	21	-3
54.5	-2	3 [12]	6 [12]	6 [0]	2 [-4]			17	-34	68	20
44.5	-3	3 [18]	5 [15]	4 [0]				12	-36	108	33
f_X		7	15	25	23	20	10	$\sum f_X = \sum f_Y = N = 100$	$\sum f_Y u_Y = -55$	$\sum f_Y u_Y^2 = 253$	$\sum f u_X u_Y = 125$
$f_X u_X$		-14	-15	0	23	40	30	$\sum f_X u_X = 64$			
$f_X u_X^2$		28	15	0	23	80	90	$\sum f_X u_X^2 = 236$			
Suma de los números en las esquinas de cada columna		32	31	0	-1	24	39	$\sum f u_X u_Y = 125$			

Comprobación

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{160.20}{(13.966)(14.925)} = 0.7686$$

que coincide con el valor obtenido en el problema 14.19.

RECTAS DE REGRESIÓN Y EL COEFICIENTE DE CORRELACIÓN

14.21 Probar que las rectas de regresión de Y sobre X y de X sobre Y son, respectivamente, *a*) $Y - \bar{Y} = (rs_Y/s_X)(X - \bar{Y})$ y *b*) $X - \bar{X} = (rs_X/s_Y)(Y - \bar{Y})$.

SOLUCIÓN

a) De acuerdo con el problema 13.15*a*), la ecuación de la recta de regresión de Y sobre X es

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \qquad \text{o bien} \qquad Y - \bar{Y} = \left(\frac{\sum xy}{\sum x^2}\right)(X - \bar{X})$$

Entonces, como $$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$ (ver problema 14.10)

se tiene $$\frac{\sum xy}{\sum x^2} = \frac{r\sqrt{(\sum x^2)(\sum y^2)}}{\sum x^2} = \frac{r\sqrt{\sum y^2}}{\sqrt{\sum x^2}} = \frac{rs_Y}{s_X}$$

de donde resulta la fórmula buscada.

b) La fórmula buscada se obtiene intercambiando X y Y en el inciso *a*).

14.22 Si las rectas de regresión de Y sobre X y de X sobre Y son, respectivamente, $Y = a_0 + a_1X$ y $X = b_0 + b_1Y$, probar que $a_1b_1 = r^2$.

SOLUCIÓN

De acuerdo con el problema 14.21, incisos *a*) y *b*),

$$a_1 = \frac{rs_Y}{s_X} \quad \text{y} \quad b_1 = \frac{rs_X}{s_Y}$$

Por lo tanto, $$a_1b_1 = \left(\frac{rs_Y}{s_X}\right)\left(\frac{rs_X}{s_Y}\right) = r^2$$

Esta fórmula puede tomarse como el punto de partida para la definición del coeficiente de correlación lineal.

14.23 Emplear la fórmula obtenida en el problema 14.22 para hallar el coeficiente de correlación lineal correspondiente a los datos del problema 14.1.

SOLUCIÓN

De acuerdo con el problema 14.1 [incisos *b*) y *c*), respectivamente] $a_1 = 484/1\,016 = 0.476$ y $b_1 = 484/467 = 1.036$. Por lo tanto, $Y^2 = a_1b_1 = (384/1\,016)(484/467)$ y $r = 0.7027$.

14.24 Dados los datos del problema 14.19, escribir las ecuaciones de las rectas de regresión de: *a*) Y sobre X y *b*) X sobre Y.

SOLUCIÓN

De acuerdo con la tabla de correlación (tabla 14.16) del problema 14.19, se tiene

$$\bar{X} = A + c_X\frac{\sum f_X u_X}{N} = 64.5 + \frac{(10)(64)}{100} = 70.9$$

$$\bar{Y} = B + c_Y\frac{\sum f_Y u_Y}{N} = 74.5 + \frac{(10)(-55)}{100} = 69.0$$

De acuerdo con los resultados del problema 14.20, $s_X = 13.966$, $s_Y = 14.925$ y $r = 0.7686$. Ahora, empleando el problema 14.21, incisos *a*) y *b*), se obtienen las ecuaciones de las rectas de regresión.

a) $Y - \bar{Y} = \frac{rs_Y}{s_X}(X - \bar{X})$ $\qquad Y - 69.0 = \frac{(0.7686)(14.925)}{13.966}(X - 70.9) = 0.821(X - 70.9)$

b) $X - \bar{X} = \frac{rs_X}{s_Y}(Y - \bar{Y})$ $\qquad X - 70.9 = \frac{(0.7686)(13.966)}{14.925}(Y - 69.0) = 0.719(Y - 69.0)$

14.25 Dados los datos del problema 14.19, calcular los errores estándar de estimación: *a*) $s_{Y.X}$ y *b*) $s_{X.Y}$. Usar los resultados del problema 14.20.

SOLUCIÓN

a) $s_{Y.X} = s_Y\sqrt{1 - r^2} = 14.925\sqrt{1 - (0.7686)^2} = 9.548$

b) $s_{X.Y} = s_X\sqrt{1 - r^2} = 13.966\sqrt{1 - (0.7686)^2} = 8.934$

14.26 En la tabla 14.17 se presentan los índices de precios al consumidor para alimentos y atención médica, de Estados Unidos, desde 2000 hasta 2006, comparados con los precios de los años base, 1982 a 1984 (tomando la media como 100). Calcular el coeficiente de correlación entre estos dos índices de precios y dar el cálculo de este coeficiente empleando MINITAB.

Tabla 14.17

Año	2000	2001	2002	2003	2004	2005	2006
Alimentos	167.8	173.1	176.2	180.0	186.2	190.7	195.2
Medicamentos	260.8	272.8	285.6	297.1	310.1	323.2	336.2

Fuente: Bureau of Labor Statistics.

SOLUCIÓN

Estos índices para alimentos y para atención médica se denotan X y Y, respectivamente, y los cálculos del coeficiente de correlación se organizan en la tabla 14.18. (Obsérvese que el año se usa únicamente para especificar los valores correspondientes a X y Y.)

Tabla 14.18

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy	y^2
167.8	260.8	−13.5	−37.2	182.25	502.20	1 383.84
173.1	272.8	−8.2	−25.2	67.24	206.64	635.04
176.2	285.6	−5.1	−12.4	26.01	63.24	153.76
180.0	297.1	−1.3	−0.9	1.69	1.17	0.81
186.2	310.1	4.9	12.1	24.01	59.29	46.41
190.7	323.2	9.4	25.2	88.36	236.88	635.04
195.2	336.2	13.9	38.2	193.21	530.98	1 459.24
$\bar{X} = 181.3$	$\bar{Y} = 298.0$			Suma = 582.77	Suma = 1 600.4	Suma = 4 414.14

Entonces, mediante la fórmula del producto-momento

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{1\,600.4}{\sqrt{(582.77)(4\,414.14)}} = 0.998$$

Después de ingresar los valores de X en `C1` y los valores de Y en `C2`, con el comando de MINITAB **`correlation C1 C2`**, se obtiene el coeficiente de correlación, que es igual al calculado antes.

```
Correlations: X, Y
Pearson correlation of X and Y =0.998
P-Value=0.000
```

CORRELACIÓN NO LINEAL

14.27 Ajustar una parábola de mínimos cuadrados de la forma $Y = a_0 + a_1X + a_2X^2$ al conjunto de datos de la tabla 14.19. Dar también la solución empleando MINITAB.

SOLUCIÓN

Las ecuaciones normales (*23*) del capítulo 13 son

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X + a_2 \sum X^2 \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum X^3 \\ \sum X^2 Y &= a_0 \sum X^2 + a_1 \sum X^3 + a_2 \sum X^4 \end{aligned} \tag{36}$$

Tabla 14.19

X	1.2	1.8	3.1	4.9	5.7	7.1	8.6	9.8
Y	4.5	5.9	7.0	7.8	7.2	6.8	4.5	2.7

Para facilitar el cálculo de las sumas se elabora la tabla 14.20. Entonces, como $N = 8$, las ecuaciones normales (*36*) se convierten en

$$\begin{aligned} 8a_0 + 42.2a_1 + 291.20a_2 &= 46.4 \\ 42.2a_0 + 291.20a_1 + 2\,275.35a_2 &= 230.42 \\ 291.20a_0 + 2\,275.35a_1 + 18\,971.92a_2 &= 1\,449.00 \end{aligned} \tag{37}$$

Resolviendo, $a_0 = 2.588$, $a_1 = 2.056$ y $a_2 = -0.2110$, de manera que la ecuación de la parábola de mínimos cuadrados buscada es

$$Y = 2.588 + 2.065X - 0.2110X^2$$

Tabla 14.20

X	Y	X^2	X^3	X^4	XY	X^2Y
1.2	4.5	1.44	1.73	2.08	5.40	6.48
1.8	5.9	3.24	5.83	10.49	10.62	19.12
3.1	7.0	9.61	29.79	92.35	21.70	67.27
4.9	7.8	24.01	117.65	576.48	38.22	187.28
5.7	7.2	32.49	185.19	1 055.58	41.04	233.93
7.1	6.8	50.41	357.91	2 541.16	48.28	342.79
8.6	4.5	73.96	636.06	5 470.12	38.70	332.82
9.8	2.7	96.04	941.19	9 223.66	26.46	259.31
$\sum X$ = 42.2	$\sum Y$ = 46.4	$\sum X^2$ = 291.20	$\sum X^3$ = 2 275.35	$\sum X^4$ = 18 971.92	$\sum XY$ = 230.42	$\sum X^2Y$ = 1 449.00

Los valores de Y se ingresan en `C1`, los valores de X se ingresan en `C2`, y los valores de X^2 se ingresan en `C3`. Se da la secuencia **Stat** → **Regresión** → **Regression** de MINITAB. La parábola de mínimos cuadrados dada como parte de los resultados es la siguiente:

```
La ecuación de regresión es Y = 2.59 + 2.06 X - 0.211 Xcuadrada
```

Que es la misma ecuación obtenida resolviendo las ecuaciones normales.

14.28 Usar la parábola de mínimos cuadrados del problema 14.27 para estimar el valor de Y para los valores dados de X.

SOLUCIÓN

Para $X = 1.2$, $Y_{est} = 2.588 + 2.065(1.2) - 0.2110(1.2)^2 = 4.762$. Los demás valores estimados se obtienen de manera similar. Los resultados se muestran en la tabla 14.21 junto con los valores reales de Y.

Tabla 14.21

Y_{est}	4.762	5.621	6.962	7.640	7.503	6.613	4.741	2.561
Y	4.5	5.9	7.0	7.8	7.2	6.8	4.5	2.7

14.29 *a*) Encontrar el coeficiente de correlación lineal entre las variables X y Y del problema 14.27.

b) Encontrar el coeficiente de correlación no lineal entre las variables X y Y del problema 14.27, asumiendo la relación parábolica obtenida en el problema 14.27.

c) Explicar la diferencia entre los coeficientes de correlación obtenidos en los incisos *a*) y *b*).

d) ¿Qué porcentaje de la variación total queda no explicada si se supone que la relación entre X y Y es la relación parabólica?

SOLUCIÓN

a) Empleando los cálculos ya realizados en la tabla 14.20 y el hecho de que $\sum Y^2 = 290.52$, se encuentra que

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

$$= \frac{(8)(230.42) - (42.2)(46.4)}{\sqrt{[(8)(291.20) - (42.2)^2][(8)(290.52) - (46.4)^2]}} = -0.3743$$

b) De acuerdo con la tabla 14.20, $\bar{Y} = (\sum Y)/N = 46.4/8 = 5.80$; por lo tanto, la variación total es $\sum (Y - \bar{Y})^2 = 21.40$. De acuerdo con la tabla 14.21, la variación explicada es $\sum (Y_{est} - \bar{Y})^2 = 21.02$. Por lo tanto,

$$r^2 = \frac{\text{variación explicada}}{\text{variación total}} = \frac{21.02}{21.40} = 0.9822 \qquad \text{y} \qquad r = 0.9911 \qquad \text{o bien} \qquad 0.99$$

c) El hecho de que en el inciso *a*) la correlación lineal sea de sólo -0.3743 indica que prácticamente no hay ninguna *relación lineal* entre X y Y. Sin embargo, hay una muy buena *relación no lineal* dada por la parábola del problema 14.27, como lo indica el hecho de que en el inciso *b*) el coeficiente de correlación sea 0.99.

d)
$$\frac{\text{Variación no explicada}}{\text{Variación total}} = 1 - r^2 = 1 - 0.9822 = 0.0178$$

Por lo tanto, 1.78% de la variación total queda no explicada. Esto puede deberse a fluctuaciones aleatorias o a otras variables que no hayan sido tomadas en consideración.

14.30 Dados los datos del problema 14.27, encontrar: *a*) s_Y y *b*) $s_{Y.X}$.

SOLUCIÓN

a) De acuerdo con el problema 14.29*a*), $\sum (Y - \bar{Y})^2 = 21.40$. Por lo tanto, la desviación estándar de Y es

$$s_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}} = \sqrt{\frac{21.40}{8}} = 1.636 \qquad \text{o bien} \qquad 1.64$$

b) **Primer método**

Empleando el inciso *a*) y el problema 14.29*b*), el error estándar de estimación de Y sobre X es

$$s_{Y.X} = s_Y\sqrt{1-r^2} = 1.636\sqrt{1-(0.9911)^2} = 0.218 \qquad \text{o bien} \qquad 0.22$$

Segundo método

Usando el problema 14.29,

$$s_{Y.X} = \sqrt{\frac{\sum (Y - Y_{\text{est}})^2}{N}} = \sqrt{\frac{\text{variación no explicada}}{N}} = \sqrt{\frac{21.40 - 21.02}{8}} = 0.218 \qquad \text{o bien} \qquad 0.22$$

Tercer método

Usando el problema 14.27 y el cálculo adicional $\sum Y^2 = 290.52$, se tiene

$$s_{Y.X} = \sqrt{\frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY - a_2 \sum X^2 Y}{N}} = 0.218 \qquad \text{o bien} \qquad 0.22$$

TEORÍA MUESTRAL DE LA CORRELACIÓN

14.31 En una muestra de tamaño 18, el coeficiente de correlación encontrado es 0.32. ¿Puede concluirse, a los niveles de significancia: *a*) 0.05 y *b*) 0.01, que el coeficiente de correlación poblacional correspondiente difiere de cero?

SOLUCIÓN

Se debe decidir entre las hipótesis $H_0 : \rho = 0$ y $H_1 : \rho > 0$.

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{0.32\sqrt{18-2}}{\sqrt{1-(0.32)^2}} = 1.35$$

a) Empleando una prueba de una cola con la distribución de Student al nivel 0.05, H_0 se rechaza si $t > t_{.95} = 1.75$ para $(18 - 2) = 16$ grados de libertad. Por lo tanto, al nivel 0.05, no se rechaza H_0.

b) Como al nivel 0.05 no se rechaza H_0, seguramente tampoco se rechazará al nivel 0.01.

14.32 ¿Cuál será el mínimo tamaño de muestra necesario para que se pueda concluir, al nivel 0.05, que un coeficiente de correlación 0.32 difiere significativamente de cero?

SOLUCIÓN

Empleando una prueba de una cola con la distribución de Student al nivel 0.05, el valor mínimo de N debe ser tal que

$$\frac{0.32\sqrt{N-2}}{\sqrt{1-(0.32)^2}} = t_{.95}$$

para $N - 2$ grados de libertad. Para un número infinito de grados de libertad $t_{.95} = 1.64$ y por lo tanto, $N = 25.6$.

Para $N = 26$: $\nu = 24$ $t_{.95} = 1.71$ $t = 0.32\sqrt{24}/\sqrt{1-(0.32)^2} = 1.65$

Para $N = 27$: $\nu = 25$ $t_{.95} = 1.71$ $t = 0.32\sqrt{25}/\sqrt{1-(0.32)^2} = 1.69$

Para $N = 28$: $\nu = 26$ $t_{.95} = 1.71$ $t = 0.32\sqrt{26}/\sqrt{1-(0.32)^2} = 1.72$

Por lo tanto, el tamaño mínimo de la muestra es $N = 28$.

14.33 En una muestra de tamaño 24, el coeficiente de correlación encontrado es $r = 0.75$. Al nivel de significancia 0.05, ¿se puede rechazar la hipótesis de que el coeficiente de correlación poblacional sea tan pequeño como: *a*) $\rho = 0.60$ y *b*) $\rho = 0.50$?

SOLUCIÓN

a) $$Z = 1.1513 \log\left(\frac{1+0.75}{1-0.75}\right) = 0.9730 \quad \mu_Z = 1.1513 \log\left(\frac{1+0.60}{1-0.60}\right) = 0.6932$$

y $$\sigma_Z = \frac{1}{\sqrt{N-3}} = \frac{1}{\sqrt{21}} = 0.2182$$

Por lo tanto, $$z = \frac{Z - \mu_Z}{\sigma_Z} = \frac{0.9730 - 0.6932}{0.2182} = 1.28$$

Empleando la distribución normal para una prueba de una cola al nivel de significancia 0.05, la hipótesis sólo se podrá rechazar si z es mayor que 1.64. Por lo tanto, en este caso no se puede rechazar la hipótesis de que el coeficiente de correlación poblacional sea tan pequeño como 0.60.

b) Si $\rho = 0.50$, entonces $\mu_Z = 1.1513 \log 3 = 0.5493$ y $z = (0.9730 - 0.5493)/0.2182 = 1.94$. Por lo tanto, la hipótesis de que el coeficiente de correlación poblacional sea tan pequeño como $\rho = 0.50$ al nivel 0.05 puede rechazarse.

14.34 Se calcula que el coeficiente de correlación entre las calificaciones finales en física y matemáticas de un grupo de 21 estudiantes es 0.80. Encontrar límites de confianza de 95% para este coeficiente.

SOLUCIÓN

Como $r = 0.80$ y $N = 21$, los límites de confianza del 95% para μ_z están dados por

$$Z \pm 1.96\sigma_Z = 1.1513 \log\left(\frac{1+r}{1-r}\right) \pm 1.96\left(\frac{1}{\sqrt{N-3}}\right) = 1.0986 \pm 0.4620$$

Por lo tanto, μ_Z tiene el intervalo de confianza de 95% siguiente: 0.5366 a 1.5606. Ahora, si

$$\mu_Z = 1.1513 \log\left(\frac{1+\rho}{1-\rho}\right) = 0.5366 \qquad \text{entonces} \qquad \rho = 0.4904$$

y si $$\mu_Z = 1.1513 \log\left(\frac{1+\rho}{1-\rho}\right) = 1.5606 \qquad \text{entonces} \qquad \rho = 0.9155$$

Por lo tanto, los límites de confianza de 95% para ρ son 0.49 y 0.92.

14.35 A partir de dos muestras de tamaño $N_1 = 28$ y $N_2 = 35$ se obtuvieron los coeficientes de correlación $r_1 = 0.50$ y $r_2 = 0.30$, respectivamente. Al nivel de significancia 0.05, ¿existe una diferencia significativa entre estos dos coeficientes?

SOLUCIÓN

$$Z_1 = 1.1513 \log\left(\frac{1+r_1}{1-r_1}\right) = 0.5493 \qquad Z_2 = 1.1513 \log\left(\frac{1+r_2}{1-r_2}\right) = 0.3095$$

y $$\sigma_{Z_1 - Z_2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} = 0.2669$$

Se debe decidir entre las hipótesis $H_0 : \mu_{Z_1} = \mu_{Z_2}$ y $H_1 : \mu_{Z_1} \neq \mu_{Z_2}$. Bajo la hipótesis H_0,

$$z = \frac{Z_1 - Z_2 - (\mu_{Z_1} - \mu_{Z_2})}{\sigma_{Z_1 - Z_2}} = \frac{0.5493 - 0.3095 - 0}{0.2669} = 0.8985$$

Empleando la distribución normal para una prueba de dos colas, H_0 se rechazará sólo si $z > 1.96$ o $z < -1.96$. Por lo tanto, no se puede rechazar H_0, y se concluye que al nivel de significancia 0.05 los resultados no son notablemente diferentes.

TEORÍA MUESTRAL DE LA REGRESIÓN

14.36 En el problema 14.1 se encontró que la ecuación de regresión de Y sobre X era $Y = 35.82 + 0.476X$. Al nivel de significancia 0.05, probar la hipótesis nula de que el coeficiente de regresión de la ecuación de regresión poblacional es 0.180 contra la hipótesis alternativa de que este coeficiente de regresión es mayor a 0.180. Realizar esta prueba sin ayuda de un software para estadística, así como con la ayuda de MINITAB.

SOLUCIÓN

$$t = \frac{a_1 - A_1}{S_{Y.X}/S_X}\sqrt{N-2} = \frac{0.476 - 0.180}{1.28/2.66}\sqrt{12-2} = 1.95$$

como $S_{Y.X} = 1.28$ (calculado en el problema 14.5) y $S_X = \sqrt{(\sum x^2)/N} = \sqrt{84.68/12} = 2.66$. Empleando una prueba de una cola con la distribución de Student al nivel 0.05, la hipótesis de que el coeficiente de regresión es 0.180 se rechazará si $t > t_{95} = 1.81$ para $(12 - 2) = 10$ grados de libertad. Por lo tanto, se rechaza la hipótesis nula.

Los resultados de MINITAB para este problema son los siguientes:

```
MTB > Regress 'Y' 1 'X';
SUBC> Constant;
SUBC> Predict C7.
```

Análisis de regresión

```
La ecuación de regresión es
Y = 35.8 + 0.476 X

Predictor        Coef        StDev          T          P
Constant        35.82        10.18       3.52      0.006
X              0.4764       0.1525       3.12      0.011

S = 1.404    R-Sq = 49.4%    R-Sq(adj) = 44.3%

Análisis de varianza

Source            DF           SS          MS         F          P
Regression         1       19.214      19.214      9.75      0.011
Residual Error    10       19.703       1.970
Total             11       38.917

Predicted Values
Fit       StDev Fit        95.0% CI            95.0% PI
66.789    0.478       (65.724,  67.855)   (63.485,  70.094)
69.171    0.650       (67.723,  70.620)   (65.724,  72.618)
```

El siguiente fragmento de los resultados proporciona la información necesaria para realizar la prueba de hipótesis.

```
Predictor        Coef        StDev          T          P
Constant        35.82        10.18       3.52      0.006
X              0.4764       0.1525       3.12      0.011
```

El estadístico de prueba calculado se encuentra como sigue:

$$t = \frac{0.4764 - 0.180}{0.1525} = 1.94$$

El valor calculado para t, **3.12**, que se muestra en los resultados de MINITAB, sirve para probar la hipótesis nula de que el coeficiente de regresión es 0. Para probar cualquier otro valor del coeficiente de regresión se necesita hacer un cálculo

como el anterior. Para probar que el coeficiente de regresión es 0.25, por ejemplo, el valor calculado para el estadístico de prueba será igual a

$$t = \frac{0.4764 - 0.25}{0.1525} = 1.48$$

La hipótesis nula de que el coeficiente de regresión es igual a 0.25 no se rechazará.

14.37 Encontrar los límites de confianza de 95% para el coeficiente de regresión del problema 14.36. Establecer el intervalo de confianza sin ayuda de un software para estadística, así como con ayuda de MINITAB.

SOLUCIÓN

El intervalo de confianza puede expresarse como

$$a_1 \pm \frac{t}{\sqrt{N-2}}\left(\frac{S_{Y.X}}{S_X}\right)$$

Por lo tanto, los límites de confianza de 95% para A_1 (obtenidos haciendo $t = \pm t_{.975} = \pm 2.23$ para $12 - 2 = 10$ grados de libertad) están dados por

$$a_1 \pm \frac{2.23}{\sqrt{12-2}}\left(\frac{S_{Y.X}}{S_X}\right) = 0.476 \pm \frac{2.23}{\sqrt{10}}\left(\frac{1.28}{2.66}\right) = 0.476 \pm 0.340$$

Es decir, se tiene una confianza de 95% de que A_1 se encuentre entre 0.136 y 0.816.

En el siguiente fragmento de los resultados obtenidos con MINITAB para el problema 14.36 aparece la información necesaria para establecer el intervalo de confianza de 95%.

Predictor	Coef	StDev	T	P
Constant	35.82	10.18	3.52	0.006
X	**0.4764**	**0.1525**	**3.12**	0.011

El término

$$\frac{1}{\sqrt{N-2}}\left(\frac{S_{Y.X}}{S_X}\right)$$

se conoce como el error estándar correspondiente al coeficiente de regresión estimado. En los resultados de MINITAB este error estándar es **0.1525**. Para hallar el intervalo de confianza de 95%, se multiplica este error estándar por $t_{.975}$, y después este término se suma y se resta a $a_1 = 0.476$, con lo que se obtiene el siguiente intervalo de confianza para A_1:

$$0.476 \pm 2.23(0.1525) = 0.476 \pm 0.340$$

14.38 En el problema 14.1, encontrar los límites de confianza de 95% para las estaturas de los hijos cuyos padres tienen una estatura de: *a*) 65.0 y *b*) 70.0 in. Encontrar el intervalo de confianza sin ayuda de software, así como con ayuda de MINITAB.

SOLUCIÓN

Como $t_{.975} = 2.23$ para $(12 - 2) = 10$ grados de libertad, los límites de confianza de 95% para Y_P están dados por

$$Y_0 \pm \frac{2.23}{\sqrt{N-2}}\, S_{Y.X}\sqrt{N + 1 + \frac{(X_0 - \bar{X})^2}{S_X^2}}$$

donde $Y_0 = 35.82 + 0.476X_0$, $S_{Y.X} = 1.28$, $S_X = 2.66$ y $N = 12$.

a) Si $X_0 = 65.0$, entonces $Y_0 = 66.76$ in. Además, $(X_0 - \bar{X})^2 = (65.0 - 66.67)^2 = 2.78$. De manera que los límites de confianza de 95% son

$$66.76 \pm \frac{2.23}{\sqrt{10}}(1.28)\sqrt{12 + 1 + \frac{2.78}{2.66^2}} = 66.76 \pm 3.30 \text{ in}$$

Es decir, se puede tener una confianza de 95% de que las estaturas de los hijos están entre 63.46 y 70.06 in.

b) Si $X_0 = 70.0$, entonces $Y_0 = 69.14$ in. Además, $(X_0 - \bar{X})^2 = (70.0 - 66.67)^2 = 11.09$. De manera que los límites de confianza de 95% son 69.14 ± 3.45 in; es decir, se puede tener una confianza de 95% de que las estaturas de los hijos estén entre 65.69 y 72.59 in.

En el siguiente fragmento de los resultados obtenidos con MINITAB para el problema 14.36 aparecen los límites de confianza para las estaturas de los hijos.

```
Predicted Values
Fit       StDev Fit        95.0% CI            95.0% PI
66.789        0.478   (65.724,  67.855)   (63.485,  70.094)
69.171        0.650   (67.723,  70.620)   (65.724,  72.618)
```

A los intervalos de confianza para individuos se les conoce como intervalos de predicción. Los intervalos de predicción del 95% aparecen en negritas. Estos intervalos coinciden con los antes calculados, salvo errores de redondeo.

14.39 En el problema 14.1, encontrar los límites de confianza de 95% para la estatura media de los hijos cuyos padres tienen una estatura de: *a*) 65.0 y *b*) 70.0 in. Establecer el intervalo de confianza sin ayuda de software, así como con ayuda de MINITAB.

SOLUCIÓN

Como $t_{.975} = 2.23$ para 10 grados de libertad, los límites de confianza de 95% para $\bar{Y}_P$ están dados por

$$Y_0 \pm \frac{2.23}{\sqrt{10}} S_{Y.X} \sqrt{1 + \frac{(X_0 - \bar{X})^2}{S_X^2}}$$

donde $Y_0 = 35.82 + 0.476X_0$, $S_{Y.X} = 1.28$, $S_X = 2.66$.

a) Para $X_0 = 65.0$, los límites de confianza serán 66.76 ± 1.07 o bien 65.7 y 67.8.

b) Para $X_0 = 70.0$, los límites de confianza serán 69.14 ± 1.45 o bien 67.7 y 70.6.

En el siguiente fragmento de los resultados obtenidos con MINITAB para el problema 14.36 aparecen los límites de confianza para las estaturas medias.

```
Predicted Values
Fit       StDev Fit        95.0% CI            95.0% PI
66.789        0.478   (65.724,  67.855)   (63.485,  70.094)
69.171        0.650   (67.723,  70.620)   (65.724,  72.618)
```

PROBLEMAS SUPLEMENTARIOS

REGRESIÓN LINEAL Y CORRELACIÓN

14.40 En la tabla 14.22 se presentan las calificaciones (denotadas *X* y *Y*, respectivamente) de 10 estudiantes en dos primeros exámenes de biología.

a) Construir un diagrama de dispersión.

b) Encontrar la recta de regresión de mínimos cuadrados de *Y* sobre *X*.

c) Encontrar la recta de regresión de mínimos cuadrados de *X* sobre *Y*.

d) Graficar las dos de rectas de regresión de los incisos *b*) y *c*) en el diagrama de dispersión del inciso *a*).

14.41 Dados los datos de la tabla 14.22, encontrar: *a*) $s_{Y.X}$ y *b*) $s_{X.Y}$.

Tabla 14.22

Calificación en el primer examen (X)	6	5	8	8	7	6	10	4	9	7
Calificación en el segundo examen (Y)	8	7	7	10	5	8	10	6	8	6

14.42 Dados los datos del problema 14.40, calcular: *a*) la variación total de Y, *b*) la variación no explicada de Y y *c*) la variación explicada de Y.

14.43 Empleando los resultados del problema 14.42, encontrar el coeficiente de correlación entre los dos conjuntos de calificaciones del problema 14.40.

14.44 Empleando la fórmula del producto-momento encontrar el coeficiente de correlación entre los dos conjuntos de calificaciones del problema 14.40; comparar el resultado con el coeficiente de correlación dado por SPSS, SAS, STATISTIX, MINITAB y EXCEL.

14.45 Dados los datos del problema 14.40*a*), encontrar la covarianza: *a*) directamente y *b*) usando la fórmula $s_{XY} = rs_X s_Y$ y los resultados de los problemas 14.43 y 14.44.

14.46 En la tabla 14.23 se presenta la edad X y la presión sistólica Y de 12 mujeres.

a) Encontrar el coeficiente de correlación entre X y Y empleando la fórmula del producto-momento, EXCEL, MINITAB, SAS, SPSS y STATISTIX.

b) Determinar la ecuación de regresión por mínimos cuadrados de Y sobre X resolviendo las ecuaciones normales y empleando EXCEL, MINITAB, SAS, SPSS y STATISTIX.

c) Estimar la presión sanguínea de una mujer de 45 años de edad.

Tabla 14.23

Edad (X)	56	42	72	36	63	47	55	49	38	42	68	60
Presión sanguínea (Y)	147	125	160	118	149	128	150	145	115	140	152	155

14.47 Encontrar los coeficientes de correlación para los datos: *a*) del problema 13.32 y *b*) del problema 13.35.

14.48 El coeficiente de correlación entre dos variables X y Y es $r = 0.60$. Si $s_X = 1.50$, $s_Y = 2.00$, $\bar{X} = 10$ y $\bar{Y} = 20$, hallar la ecuación de la recta de regresión: *a*) de Y sobre X y *b*) de X sobre Y.

14.49 Dados los datos del problema 14.48, calcular: *a*) $s_{Y.X}$ y *b*) $s_{X.Y}$.

14.50 Si $s_{Y.X} = 3$ y $s_Y = 5$, hallar r.

14.51 Si el coeficiente de correlación entre X y Y es 0.50, ¿qué porcentaje de la variación total queda no explicada por la ecuación de regresión?

14.52 *a*) Probar que la ecuación de la recta de regresión de Y sobre X puede expresarse como

$$Y - \bar{Y} = \frac{s_{XY}}{s_X^2}(X - \bar{X})$$

b) Escribir la ecuación análoga para la recta de regresión de X sobre Y.

14.53 *a*) Calcular el coeficiente de correlación entre los valores correspondientes de X y Y dados en la tabla 14.24.

Tabla 14.24

X	2	4	5	6	8	11
Y	18	12	10	8	7	5

b) Multiplicar por 2 cada uno de los valores de X que aparecen en la tabla y sumarles 6. Multiplicar por 3 cada uno de los valores de Y que aparecen en la tabla y restarles 15. Encontrar el coeficiente de correlación entre estos dos nuevos conjuntos de valores y explicar por qué sí, o por qué no, se obtienen los mismos resultados que en el inciso *a*).

14.54 *a*) Dados los datos del problema 14.53, incisos *a*) y *b*), encontrar las ecuaciones de regresión de Y sobre X.

b) Analizar la relación entre estas dos ecuaciones de regresión.

14.55 *a*) Probar que el coeficiente de correlación entre X y Y se puede expresar como

$$r = \frac{\overline{XY} - \bar{X}\bar{Y}}{\sqrt{[\overline{X^2} - \bar{X}^2][\overline{Y^2} - \bar{Y}^2]}}$$

b) Aplicar este método al problema 14.1.

14.56 Probar que el coeficiente de correlación es independiente de la elección del origen de las variables o de las unidades en las que estén expresadas. (*Sugerencia:* Suponga que $X' = c_1X + A$ y $Y' = c_2Y + B$ donde c_1, c_2, A y B son constantes cualesquiera, y probar que el coeficiente de correlación entre X' y Y' es el mismo que entre X y Y.)

14.57 *a*) Probar que, para la regresión lineal,

$$\frac{s_{Y.X}^2}{s_Y^2} = \frac{s_{X.Y}^2}{s_X^2}$$

b) ¿Es válido este resultado para la regresión no lineal?

COEFICIENTE DE CORRELACIÓN PARA DATOS AGRUPADOS

14.58 Encontrar el coeficiente de correlación entre las estaturas y los pesos de 300 hombres adultos, presentadas en la tabla 14.25, una tabla de frecuencias.

Tabla 14.25

		Estaturas X (in)				
		59-62	63-66	67-70	71-74	75-78
Pesos Y (lb)	90-109	2	1			
	110-129	7	8	4	2	
	130-149	5	15	22	7	1
	150-169	2	12	63	19	5
	170-189		7	28	32	12
	190-209		2	10	20	7
	210-229			1	4	2

14.59 *a*) Dados los datos del problema 14.58, encontrar la ecuación de regresión por mínimos cuadrados de Y sobre X.
b) Estimar los pesos de los hombres cuyas estaturas son 64 y 72 in, respectivamente.

14.60 Dados los datos del problema 14.58, encontrar: *a*) $s_{Y.X}$ y *b*) $s_{X.Y}$.

14.61 Establecer la fórmula (*21*) de este capítulo para el coeficiente de correlación de datos agrupados.

CORRELACIÓN DE SERIES DE TIEMPO

14.62 En la tabla 14.26 se presenta el gasto anual promedio, por consumidor, en atención a la salud y el ingreso per cápita desde 1999 hasta 2004. Encontrar el coeficiente de correlación.

Tabla 14.26

Año	1999	2000	2001	2002	2003	2004
Costo de la atención a la salud	1 959	2 066	2 182	2 350	2 416	2 574
Ingreso per cápita	27 939	29 845	30 574	30 810	31 484	33 050

Fuente: Bureau of Laboral Statistics and U.S. Bureau of Economic Analysis.

14.63 En la tabla 14.27 se muestran temperatura y precipitación promedio durante el mes de julio en una ciudad, desde 2000 hasta 2006. Hallar el coeficiente de correlación.

Tabla 14.27

Año	2000	2001	2002	2003	2004	2005	2006
Temperatura (°F)	78.1	71.8	75.6	72.7	75.3	73.6	75.1
Precipitación (in)	6.23	3.64	3.42	2.84	1.83	2.82	4.04

TEORÍA MUESTRAL DE LA CORRELACIÓN

14.64 En una muestra de tamaño 27, el coeficiente de correlación calculado es 0.40. ¿Puede concluirse a los niveles de significancia: *a*) 0.05 y *b*) 0.01, que el coeficiente de correlación poblacional correspondiente sea distinto de cero?

14.65 En una muestra de tamaño 35, el coeficiente de correlación calculado es 0.50. ¿Puede concluirse al nivel de significancia 0.05 que el coeficiente de correlación poblacional sea: *a*) tan pequeño como $\rho = 0.30$ y *b*) tan grande como $\rho = 0.70$?

14.66 Encontrar los límites de confianza de: *a*) 95% y *b*) 99% para un coeficiente de correlación que se ha calculado que es 0.60 a partir de una muestra de tamaño 28.

14.67 Resolver el problema 14.66 si la muestra es de tamaño 52.

14.68 Encontrar los límites de confianza de 95% para el coeficiente de correlación calculado: *a*) en el problema 14.46 y *b*) en el problema 14.58.

14.69 Los coeficientes de correlación obtenidos a partir de dos muestras, una de tamaño 23 y otra de tamaño 28, fueron 0.80 y 0.95, respectivamente. ¿Puede concluirse a los niveles de significancia: *a*) 0.05 y *b*) 0.01, que existe una diferencia significativa entre estos dos coeficientes?

TEORÍA MUESTRAL DE LA REGRESIÓN

14.70 Basándose en una muestra de tamaño 27, la ecuación de regresión de Y sobre X encontrada es $Y = 25.0 + 2.00X$. Si $s_{Y.X} = 1.50$, $s_{Y.X} = 1.50$, $s_X = 3.00$ y $\bar{X} = 7.50$, encontrar los límites de confianza de *a*) 95% y *b*) 99% para el coeficiente de regresión.

14.71 Dados los datos del problema 14.70, al nivel de significancia 0.01, probar la hipótesis de que el coeficiente de regresión poblacional es: *a*) tan bajo como 1.70 y *b*) tan alto como 2.20.

14.72 Dados los datos del problema 14.70, encontrar los límites de confianza: *a*) de 95% y *b*) de 99% para Y cuando $X = 6.00$.

14.73 Dados los datos del problema 14.70, encontrar los límites de confianza: *a*) de 95% y *b*) de 99% para la media de todos los valores de Y correspondientes a $X = 6.00$.

14.74 Dados los datos del problema 14.46, encontrar los límites de confianza de 95% para: *a*) el coeficiente de regresión de Y sobre X, *b*) las presiones sanguíneas de todas las mujeres cuya edad es de 45 años y *c*) la media de las presiones sanguíneas de todas las mujeres de 45 años.

CORRELACIÓN MÚLTIPLE Y CORRELACIÓN PARCIAL

15

CORRELACIÓN MÚLTIPLE

Al grado de relación que existe entre tres o más variables se le conoce como *correlación múltiple*. Los principios fundamentales relacionados con los problemas de correlación múltiple son análogos a los de los problemas de correlación simple, tratados en el capítulo 14.

NOTACIÓN EMPLEANDO SUBÍNDICES

Para generalizar a un número mayor de variables conviene adoptar una notación con subíndices.

Sean $X_1, X_2, X_3, \ldots$ las variables en consideración. Entonces, con $X_{11}, X_{12}, X_{13}, \ldots$ se denotan los valores que asume la variable X_1, y $X_{21}, X_{22}, X_{23}, \ldots$ denotan los valores que asume la variable X_2, y así sucesivamente. Con esta notación, una suma como $X_{21} + X_{22} + X_{23} + \cdots + X_{2N}$ puede expresarse como $\sum_{j=1}^{N} X_{2j}$, $\sum_j X_{2j}$ o simplemente $\sum X_2$. Cuando no puede haber lugar a ambigüedad, se usa la última notación. En este caso, la media de X_2 se expresa: $\bar{X}_2 = \sum X_2/N$.

ECUACIONES DE REGRESIÓN Y PLANOS DE REGRESIÓN

Una *ecuación de regresión* es una ecuación que se utiliza para estimar una variable dependiente, por ejemplo X_1, a partir de las variables independientes $X_2, X_3, \ldots$ y se le llama *ecuación de regresión de X_1 sobre $X_2, X_3, \ldots$*. Empleando la notación funcional esto puede expresarse brevemente como $X_1 = F(X_2, X_3, \ldots)$ (que se lee "X_1 es una función de X_2, X_3, etcétera").

En el caso de tres variables, la ecuación de regresión más simple de X_1 sobre X_2 y X_3 tiene la forma siguiente:

$$X_1 = b_{1.23} + b_{12.3}X_2 + b_{13.2}X_3 \tag{1}$$

donde $b_{1.23}$, $b_{12.3}$ y $b_{13.2}$ son constantes. Si en la ecuación (*1*) X_3 se mantiene constante, la gráfica de X_1 *versus* X_2 es una línea recta cuya pendiente es $b_{12.3}$. Si X_2 se mantiene constante, la gráfica de X_1 *versus* X_3 es una línea recta cuya pendiente es $b_{13.2}$. Como se ve, el subíndice después del punto indica la variable que se mantiene constante en cada caso.

Dado que X_1 varía parcialmente debido a la variación de X_2 y parcialmente debido a la variación de X_3, a $b_{12.3}$ y $b_{13.2}$ se les llama *coeficientes de regresión parcial* de X_1 sobre X_2 manteniendo X_3 constante y de X_1 sobre X_3 manteniendo X_2 constante, respectivamente.

A la ecuación (*1*) se le llama *ecuación de regresión lineal* de X_1 sobre X_2 y X_3. En un sistema rectangular tridimensional de coordenadas, esta ecuación representa un plano al que se le conoce como *plano de regresión*, que es una generalización de la recta de regresión para dos variables, considerada en el capítulo 13.

ECUACIONES NORMALES PARA LOS PLANOS DE REGRESIÓN DE MÍNIMOS CUADRADOS

Así como existen rectas de regresión de mínimos cuadrados que aproximan un conjunto de puntos (X, Y) en un diagrama de dispersión bidimensional, también existen *planos de regresión de mínimos cuadrados* que se ajustan a un conjunto de N puntos (X_1, X_2, X_3) en un diagrama de dispersión tridimensional.

El plano de regresión de mínimos cuadrados de X_1 sobre X_2 y X_3 tiene la ecuación (*1*), donde $b_{1.23}$, $b_{12.3}$ y $b_{13.2}$ se determinan resolviendo simultáneamente las *ecuaciones normales*

$$\begin{aligned} \sum X_1 &= b_{1.23}N + b_{12.3}\sum X_2 + b_{13.2}\sum X_3 \\ \sum X_1X_2 &= b_{1.23}\sum X_2 + b_{12.3}\sum X_2^2 + b_{13.2}\sum X_2X_3 \\ \sum X_1X_3 &= b_{1.23}\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum X_3^2 \end{aligned} \tag{2}$$

Estas ecuaciones pueden obtenerse formalmente multiplicando, en cada caso, ambos lados de la ecuación (*1*) por 1, por X_2 y por X_3, y sumando después ambos lados.

A menos que se especifique otra cosa, siempre que se haga referencia a una ecuación de regresión se entenderá que se está haciendo referencia a la ecuación de regresión de mínimos cuadrados.

Si $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$ y $x_3 = X_3 - \bar{X}_3$, la ecuación de regresión de X_1 sobre X_2 y X_3 puede expresarse de manera más sencilla como

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \tag{3}$$

donde $b_{12.3}$ y $b_{13.2}$ se obtienen resolviendo simultáneamente las ecuaciones

$$\begin{aligned} \sum x_1x_2 &= b_{12.3}\sum x_2^2 + b_{13.2}\sum x_2x_3 \\ \sum x_1x_3 &= b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2 \end{aligned} \tag{4}$$

Estas ecuaciones, que son equivalentes a las ecuaciones normales (*2*), se obtienen formalmente multiplicando, de manera sucesiva, ambos lados de la ecuación (*3*) por x_2 y por x_3, y después sumando ambos lados (ver problema 15.8).

PLANOS DE REGRESIÓN Y COEFICIENTES DE CORRELACIÓN

Si los coeficientes de correlación entre las variables X_1 y X_2, X_1 y X_3, y X_2 y X_3, que se calcularon en el capítulo 14, se denotan respectivamente r_{12}, r_{13} y r_{23} (también llamados *coeficientes de correlación de orden cero*), entonces la ecuación del plano de regresión de mínimos cuadrados tiene la ecuación

$$\frac{x_1}{s_1} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\frac{x_2}{s_2} + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\frac{x_3}{s_3} \tag{5}$$

donde $x_1 = X - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$ y $x_3 = X_3 - \bar{X}_3$, y donde s_1, s_2 y s_3 son, respectivamente, las desviaciones estándar de X_1, X_2 y X_3 (ver problema 15.9).

Obsérvese que si la variable X_3 no existe y si $X_1 = Y$ y $X_2 = X$ entonces la ecuación (*5*) se reduce a la ecuación (*25*) del capítulo 14.

ERROR ESTÁNDAR DE ESTIMACIÓN

Mediante una obvia generalización de la ecuación (*8*) del capítulo 14 se define el *error estándar de estimación* de X_1 sobre X_2 y X_3 como

$$s_{1.23} = \sqrt{\frac{\sum(X_1 - X_{1,\text{est}})^2}{N}} \tag{6}$$

donde $X_{1,\text{est}}$ indica los valores estimados de X_1 obtenidos con las ecuaciones de regresión (*1*) o (*5*).

El error estándar de estimación también se puede calcular en términos de los coeficientes de correlación r_{12}, r_{13} y r_{23}, empleando la fórmula

$$s_{1.23} = s_1\sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \tag{7}$$

La interpretación muestral del error estándar de estimación para dos variables, dada en la página 313 para el caso en el que N es grande, puede extenderse a tres dimensiones reemplazando las rectas paralelas a la recta de regresión por planos paralelos al plano de regresión. La fórmula $\hat{s}_{1.23} = \sqrt{N/(N-3)}s_{1.23}$ proporciona una mejor estimación del error estándar de estimación poblacional.

COEFICIENTE DE CORRELACIÓN MÚLTIPLE

El coeficiente de correlación múltiple se define mediante una extensión de la ecuación (*12*) o (*14*) del capítulo 14. En el caso de dos variables independientes, por ejemplo, el *coeficiente de correlación múltiple* está dado por

$$R_{1.23} = \sqrt{1 - \frac{s_{1.23}^2}{s_1^2}} \tag{8}$$

donde s_1 es la desviación estándar de la variable X_1, y $s_{1.23}$ está dado por la ecuación (*6*) o por la ecuación (*7*). La cantidad $R_{1.23}^2$ se conoce como *coeficiente de determinación múltiple*.

Cuando se emplea una ecuación de regresión lineal, al coeficiente de correlación múltiple se le llama *coeficiente de correlación lineal múltiple*. A menos que se especifique otra cosa, el término correlación múltiple se empleará para correlación lineal múltiple.

La ecuación (*8*) también puede expresarse en términos de r_{12}, r_{13} y r_{23} como

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \tag{9}$$

El valor de un coeficiente de correlación múltiple, como $R_{1.23}$, está entre 0 y 1, inclusive. Cuanto más cerca está de 1, mejor es la relación lineal entre las variables. Cuanto más cerca esté de 0, peor será la relación lineal entre las variables. Si un coeficiente de correlación múltiple es 1, a esa correlación se le llama correlación *perfecta*. Aunque un coeficiente de correlación sea 0, esto indica que no hay relación lineal entre las variables, pero puede que exista una *relación no lineal*.

CAMBIO DE LA VARIABLE DEPENDIENTE

Los resultados anteriores son válidos cuando X_1 se considera la variable dependiente. Pero si en lugar de X_1 quiere considerarse a X_3 (por ejemplo) como la variable dependiente, lo único que hay que hacer es sustituir, en las fórmulas

ya obtenidas, el subíndice 1 por el subíndice 3 y el subíndice 3 por el subíndice 1. Por ejemplo, la ecuación de regresión de X_3 sobre X_1 y X_2 es

$$\frac{x_3}{s_3} = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2}\right)\frac{x_2}{s_2} + \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right)\frac{x_1}{s_1} \tag{10}$$

de acuerdo con la ecuación (*5*) y empleando las igualdades $r_{32} = r_{23}$, $r_{31} = r_{13}$ y $r_{21} = r_{12}$.

GENERALIZACIONES A MÁS DE TRES VARIABLES

Estas generalizaciones se obtienen por analogía con los resultados anteriores. Por ejemplo, la ecuación de regresión lineal de X_1 sobre X_2, X_3 y X_4 se expresa

$$X_1 = b_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 \tag{11}$$

y representa un *hiperplano en el espacio de cuatro dimensiones*. Multiplicando sucesivamente ambos lados de la ecuación (*11*) por 1, X_2, X_3 y X_4 y después sumando ambos lados se obtienen las ecuaciones normales con las que se determina $b_{1.234}$, $b_{12.34}$, $b_{13.24}$ y $b_{14.23}$; sustituyendo sus valores en la ecuación (*11*) se obtiene la *ecuación de regresión de mínimos cuadrados de X_1 sobre X_2, X_3 y X_4*. Esta ecuación de regresión de mínimos cuadrados se puede expresar en forma similar a la de la ecuación (*5*). (Ver problema 15.41.)

CORRELACIÓN PARCIAL

También es importante medir la correlación entre una variable dependiente y determinada variable independiente cuando todas las demás variables permanecen constantes; es decir, cuando se eliminan los efectos de todas las demás variables. Esto se logra definiendo un *coeficiente de correlación parcial*, como la ecuación (*12*) del capítulo 14, salvo que deberán considerarse las variaciones explicadas y no explicadas que surgen con esa determinada variable independiente y sin ella.

Si $r_{12.3}$ denota el coeficiente de correlación parcial entre X_1 y X_2 cuando X_3 permanece constante, se encuentra que

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \tag{12}$$

De manera similar, si $r_{12.34}$ denota el coeficiente de correlación parcial entre X_1 y X_2 cuando X_3 y X_4 permanecen constantes, entonces

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{(1 - r_{13.4}^2)(1 - r_{23.4}^2)}} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}} \tag{13}$$

Estos resultados son útiles, pues mediante ellos puede hacerse que cualquier coeficiente de correlación parcial dependa finalmente de los coeficientes de correlación r_{12}, r_{23}, etc. (es decir, de los *coeficientes de correlación de orden cero*).

Se vio que en el caso de dos variables, X y Y, si las ecuaciones de las dos rectas de regresión son $Y = a_0 + a_1X$ y $X = b_0 + b_1Y$, se tiene que $r^2 = a_1b_1$ (ver problema 14.22). Este resultado puede generalizarse. Por ejemplo, si

$$X_1 = b_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 \tag{14}$$

y

$$X_4 = b_{4.123} + b_{41.23}X_1 + b_{42.13}X_2 + b_{43.12}X_3 \tag{15}$$

son, respectivamente, las ecuaciones de regresión lineal de X_1 sobre X_2, X_3 y X_4 y de X_4 sobre X_1, X_2 y X_3, entonces

$$r_{14.23}^2 = b_{14.23}b_{41.23} \tag{16}$$

(ver problema 15.18). Esta fórmula puede tomarse como punto de partida para una definición de los coeficientes de correlación lineal parcial.

RELACIONES ENTRE COEFICIENTES DE CORRELACIÓN MÚLTIPLE Y COEFICIENTES DE CORRELACIÓN PARCIAL

Se pueden encontrar resultados interesantes que relacionan los coeficientes de correlación múltiple. Por ejemplo, se encuentra que

$$1 - R_{1.23}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2) \tag{17}$$

$$1 - R_{1.234}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2) \tag{18}$$

Las generalizaciones de estos resultados son fáciles de efectuar.

REGRESIÓN MÚLTIPLE NO LINEAL

Los resultados anteriores para la regresión lineal múltiple se pueden extender a la regresión no lineal múltiple. Los coeficientes de correlación parcial y de correlación múltiple pueden definirse mediante métodos similares a los proporcionados antes.

PROBLEMAS RESUELTOS

ECUACIONES DE REGRESIÓN CON TRES VARIABLES

15.1 Usando la notación adecuada mediante subíndices, dar la ecuación de regresión de: *a*) X_2 sobre X_1 y X_3; *b*) X_3 sobre X_1, X_2 y X_4, y *c*) X_5 sobre X_1, X_2, X_3 y X_4.

SOLUCIÓN

a) $X_2 = b_{2.13} + b_{21.3}X_1 + b_{23.1}X_3$

b) $X_3 = b_{3.124} + b_{31.24}X_1 + b_{32.14}X_2 + b_{34.12}X_4$

c) $X_5 = b_{5.1234} + b_{51.234}X_1 + b_{52.134}X_2 + b_{53.124}X_3 + b_{54.123}X_4$

15.2 Dar las ecuaciones normales correspondientes a las ecuaciones *a*) $X_3 = b_{3.12} + b_{31.2}X_1 + b_{32.1}X_2$ y *b*) $X_1 = b_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$.

SOLUCIÓN

a) La ecuación se multiplica, sucesivamente, por 1, X_1 y X_2 y se suma a ambos lados. Las ecuaciones normales son

$$\begin{aligned}
\textstyle\sum X_3 &= b_{3.12}N + b_{31.2}X_1 + b_{32.1}\sum X_2 \\
\textstyle\sum X_1X_3 &= b_{3.12}\sum X_1 + b_{31.2}\sum X_1^2 + b_{32.1}\sum X_1X_2 \\
\textstyle\sum X_2X_3 &= b_{3.12}\sum X_2 + b_{31.2}\sum X_1X_2 + b_{32.1}\sum X_2^2
\end{aligned}$$

b) La ecuación se multiplica, sucesivamente, por 1, X_2, X_3 y X_4 y se suma a ambos lados. Las ecuaciones normales son

$$\begin{aligned}
\sum X_1 &= b_{1.234}N + b_{12.34}\sum X_2 + b_{13.24}\sum X_3 + b_{14.23}\sum X_4 \\
\sum X_1X_2 &= b_{1.234}\sum X_2 + b_{12.34}\sum X_2^2 + b_{13.24}\sum X_2X_3 + b_{14.23}\sum X_2X_4 \\
\sum X_1X_3 &= b_{1.234}\sum X_3 + b_{12.34}\sum X_2X_3 + b_{13.24}\sum X_3^2 + b_{14.23}\sum X_3X_4 \\
\sum X_1X_4 &= b_{1.234}\sum X_4 + b_{12.34}\sum X_2X_4 + b_{13.24}\sum X_3X_4 + b_{14.23}\sum X_4^2
\end{aligned}$$

Obsérvese que éstas no son deducciones de las ecuaciones normales, sino únicamente una manera formal para recordarlas.

El número de ecuaciones normales es igual al número de constantes desconocidas.

15.3 En la tabla 15.1 se presentan los pesos X_1 dados a la libra (lb) más cercana, las estaturas X_2 a la pulgada (in) más cercana y las edades X_3 al año más cercano de 12 niños.

a) Encontrar la ecuación de regresión de mínimos cuadrados de X_1 sobre X_2 y X_3.
b) Determinar los valores estimados de X_1 a partir de los valores dados de X_2 y X_3.
c) Estimar el peso de un niño de 9 años que mide 54 in.
d) Encontrar la ecuación de regresión de mínimos cuadrados empleando EXCEL, MINITAB, SPSS y STATISTIX.

Tabla 15.1

Peso (X_1)	64	71	53	67	55	58	77	57	56	51	76	68
Estatura (X_2)	57	59	49	62	51	50	55	48	52	42	61	57
Edad (X_3)	8	10	6	11	8	7	10	9	10	6	12	9

SOLUCIÓN

a) La ecuación de regresión de mínimos cuadrados de X_1 sobre X_2 y X_3 puede expresarse como

$$X_1 = b_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

Las ecuaciones normales de la ecuación de regresión de mínimos cuadrados son

$$\begin{aligned}
\sum X_1 &= b_{1.23}N + b_{12.3}\sum X_2 + b_{13.2}\sum X_3 \\
\sum X_1X_2 &= b_{1.23}\sum X_2 + b_{12.3}\sum X_2^2 + b_{13.2}\sum X_2X_3 \\
\sum X_1X_3 &= b_{1.23}\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum X_3^2
\end{aligned} \qquad (19)$$

Para calcular las sumas se elabora la tabla 15.2. (Aunque la columna con el encabezado X_1^2 no se necesita en este momento, se ha incluido para referencias futuras.) Empleando la tabla 15.2, las ecuaciones normales (*19*) se convierten en

$$\begin{aligned}
12b_{1.23} + 643b_{12.3} + 106b_{13.2} &= 753 \\
643b_{1.23} + 34\,843b_{12.3} + 5\,779b_{13.2} &= 40\,830 \\
106b_{1.23} + 5\,779b_{12.3} + 976b_{13.2} &= 6\,796
\end{aligned} \qquad (20)$$

Resolviendo, $b_{1.23} = 3.6512$, $b_{12.3} = 0.8546$ y $b_{13.2} = 1.5063$, con lo que la ecuación de regresión es

$$X_1 = 3.6512 + 0.8546X_2 + 1.5063X_3 \quad \text{o} \quad X_1 = 3.65 + 0.855X_2 + 1.506X_3 \tag{21}$$

Tabla 15.2

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3
64	57	8	4 096	3 249	64	3 648	512	456
71	59	10	5 041	3 481	100	4 189	710	590
53	49	6	2 809	2 401	36	2 597	318	294
67	62	11	4 489	3 844	121	4 154	737	682
55	51	8	3 025	2 601	64	2 805	440	408
58	50	7	3 364	2 500	49	2 900	406	350
77	55	10	5 929	3 025	100	4 235	770	550
57	48	9	3 249	2 304	81	2 736	513	432
56	52	10	3 136	2 704	100	2 912	560	520
51	42	6	2 601	1 764	36	2 142	306	252
76	61	12	5 776	3 721	144	4 636	912	732
68	57	9	4 624	3 249	81	3 876	612	513
$\sum X_1 = 753$	$\sum X_2 = 643$	$\sum X_3 = 106$	$\sum X_1^2 = 48\,139$	$\sum X_2^2 = 34\,843$	$\sum X_3^2 = 976$	$\sum X_1X_2 = 40\,830$	$\sum X_1X_3 = 6\,796$	$\sum X_2X_3 = 5\,779$

En el problema 15.6 se presenta otro método en el que se evita tener que resolver ecuaciones simultáneas.

b) Sustituyendo, en la ecuación de regresión (*21*), X_2 y X_3 por sus valores se obtienen los valores estimados para X_1, que se denotan $X_{1,\text{est}}$. Por ejemplo, sustituyendo en la ecuación (*21*) $X_2 = 57$ y $X_3 = 8$, se obtiene $X_{1,\text{est}} = 64.414$.

De manera similar se obtienen los demás valores estimados para X_1. Estos valores se dan en la tabla 15.3 junto con los valores muestrales de X_1.

Tabla 15.3

$X_{1,\text{est}}$	64.414	69.136	54.564	73.206	59.286	56.925	65.717	58.229	63.153	48.582	73.857	65.920
X_1	64	71	53	67	55	58	77	57	56	51	76	68

c) Haciendo $X_2 = 54$ y $X_3 = 9$ en la ecuación (*21*), se obtiene el peso estimado $X_{1,\text{est}} = 63.356$, o 63 lb, aproximadamente.

d) En la figura 15-1 se muestra parte de los resultados obtenidos con EXCEL. Para obtener estos resultados se emplea la secuencia **Tools** → **Data analysis** → **Regression**. En los resultados, los coeficientes $b_{1.23} = 3.6512$, $b_{12.3} = 0.8546$ y $b_{13.2} = 1.5063$ aparecen en negritas.

Parte de los resultados de MINITAB es `la ecuación de regresión X1 = 3.7 + 0.855X2 + 1.51X3`. Una vez ingresados los datos en `C1-C3` se emplea la secuencia **Stat** → **Regression** → **Regression**.

En la figura 15-2 se presenta una parte de los resultados de SPSS. Los resultados se obtienen empleando la secuencia **analyze** → **Regression** → **Linear**. En los resultados, los coeficientes $b_{1.23} = 3.651$, $b_{12.3} = 0.855$ y $b_{13.2} = 1.506$ aparecen en la columna titulada `Unstandardized Coefficients`.

En la figura 15-3 se presenta parte de los resultados de STATISTIX. Los resultados se obtienen empleando la secuencia **Stastitics** → **Linear models** → **Linear Regression**.

X1	X2	X3
64	57	8
71	59	10
53	49	6
67	62	11
55	51	8
58	50	7
77	55	10
57	48	9
56	52	10
51	42	6
76	61	12
68	57	9

RESUMEN

Estadísticos de la regresión	
R^2 múltiple	0.841757
R^2 cuadrado	0.708554
R^2 ajustado	0.643789
Error estándar	5.363215
Observaciones	12

ANÁLISIS DE VARIANZA

	df
Regresión	2
Residuos	9
Total	11

	Coeficientes
Intersección	**3.651216**
X2	**0.85461**
X3	**1.506332**

Figura 15-1 EXCEL, resultados para el problema 15.3*d*).

Coeficientes[a]

Modelo		Coeficientes desestandarizados		Coeficientes estandarizados		
		B	Error estándar	Beta	t	Sig.
1	(Constante)	3.651	16.168		.226	.826
	X2	.855	.452	.565	1.892	.091
	X3	1.506	1.414	.318	1.065	.315

[a]Variable dependiente: X1

Figura 15-2 SPSS, resultados para el problema 15.3*d*).

Statistix 8.0

Regresión lineal de mínimos cuadrados de X1 de bajo peso

Variables predichas	Coeficiente	Error estándar	T	P	VIF
Constante	**3.65122**	16.1678	0.23	0.8264	
X2	**0.85461**	0.45166	1.89	0.0910	2.8
X3	**1.50633**	1.41427	1.07	0.3146	2.8

Figura 15-3 STATISTIX, resultados para el problema 15.3*d*).

Las soluciones del software son las mismas que las de las ecuaciones normales.

15.4 Dados los datos del problema 15.3, calcular las desviaciones estándar: *a*) s_1, *b*) s_2 y *c*) s_3.

SOLUCIÓN

a) La cantidad s_1 es la desviación estándar de la variable X_1. Entonces, empleando la tabla 15.2 del problema 15.3 y los métodos del capítulo 4, se encuentra

$$s_1 = \sqrt{\frac{\sum X_1^2}{N} - \left(\frac{\sum X_1}{N}\right)^2} = \sqrt{\frac{48\,139}{12} - \left(\frac{753}{12}\right)^2} = 8.6035 \quad \text{u} \quad 8.6 \text{ lb}$$

b)
$$s_2 = \sqrt{\frac{\sum X_2^2}{N} - \left(\frac{\sum X_2}{N}\right)^2} = \sqrt{\frac{34\,843}{12} - \left(\frac{643}{12}\right)^2} = 5.6930 \quad \text{o bien} \quad 5.7 \text{ in}$$

c)
$$s_3 = \sqrt{\frac{\sum X_3^2}{N} - \left(\frac{\sum X_3}{N}\right)^2} = \sqrt{\frac{976}{12} - \left(\frac{106}{12}\right)^2} = 1.8181 \quad \text{o bien} \quad 1.8 \text{ años}$$

15.5 Dados los datos del problema 15.3, calcular: *a*) r_{12}, *b*) r_{13}, *c*) r_{23}. Calcular las tres correlaciones empleando EXCEL, MINITAB y STATISTIX.

SOLUCIÓN

a) La cantidad r_{12} es el coeficiente de correlación lineal entre las variables X_1 y X_2, ignorando a la variable X_3. Por lo tanto, empleando los métodos del capítulo 14, se tiene

$$r_{12} = \frac{N\sum X_1X_2 - (\sum X_1)(\sum X_2)}{\sqrt{[N\sum X_1^2 - (\sum X_1)^2][N\sum X_2^2 - (\sum X_2)^2]}}$$

$$= \frac{(12)(40{,}830) - (753)(643)}{\sqrt{[(12)(48{,}139) - (753)^2][(12)(34{,}843) - (643)^2]}} = 0.8196 \qquad \text{o bien} \qquad 0.82$$

b) y *c*) Empleando las fórmulas correspondientes se obtiene $r_{12} = 0.7698$, o bien 0.77, y $r_{23} = 0.7984$, o bien 0.80.

d) Usando EXCEL se tiene:

A	B	C	D	E
X1	X2	X3	0.819645	=CORREL(A2:A13,B2:B13)
64	57	8	0.769817	=CORREL(A2:A13,C2:C13)
71	59	10	0.798407	=CORREL(B2:B13,C2:C13)
53	49	6		
67	62	11		
55	51	8		
58	50	7		
77	55	10		
57	48	9		
56	52	10		
51	42	6		
76	61	12		
68	57	9		

Como se ve, r_{12} está en `D1`, r_{13} está en `D2` y r_{23} está en `D3`. En `E1`, `E2` y `E3` aparecen las funciones de EXCEL empleadas para obtener los resultados.

Usando MINITAB, la secuencia **Stat** → **Basic Statistics** → **Correlation** da el resultado siguiente.

```
Correlaciones: X1, X2, X3
          X1        X2
X2     0.820
       0.001

X3     0.770     0.798
       0.003     0.002

Cell Contents: Pearson correlation
                  P-Value
```

La correlación r_{12} está en la intersección de `X1` y `X2` y es 0.820. El valor debajo de él, 0.001, es el valor p para probar que no hay correlación poblacional entre `X1` y `X2`. Como este valor p es menor de 0.05, se rechaza la hipótesis nula de que no hay correlación poblacional entre la estatura (`X2`) y el peso (`X1`). Las demás correlaciones con sus valores p se leen de manera similar.

Empleando en SPSS la secuencia **Analyze** → **Correlate** → **Bivariate** da el siguiente resultado que se lee de manera similar al de MINITAB.

Correlaciones

		X1	X2	X3
X1	Correlación de Pearson	1	.820**	.770**
	Sig. (2 colas)		.001	.003
	N	12	12	12
X2	Correlación de Pearson	.820**	1	.798**
	Sig. (2 colas)	.001		.002
	N	12	12	12
X3	Correlación de Pearson	.770**	.798**	1
	Sig. (2 colas)	.003	.002	
	N	12	12	12

**La correlación es significativa al nivel 0.01 (2 colas).

Empleando STATISTIX, la secuencia **Stastitics** → **Linear models** → **Correlation** da el resultado siguiente, que es similar al de los otros software.

```
Statistix 8.0
Correlations (Pearson)
                    X1              X2
X2                0.8196
P-VALUE           0.0011

X3                0.7698          0.7984
                  0.0034          0.0018
```

Una vez más se ve la cantidad de tiempo que se ahorra con un software que realiza los cálculos para el usuario.

15.6 Repetir el problema 15.3*a*) empleando la ecuación (*5*) de este capítulo y los resultados de los problemas 15.4 y 15.5.

SOLUCIÓN

Multiplicando ambos lados de la ecuación (*5*) por s_1, la ecuación de regresión de X_1 sobre X_2 y X_3 es,

$$x_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_2}\right)x_2 + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_3}\right)x_3 \qquad (22)$$

donde $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$ y $x_3 = X_3 - \bar{X}_3$. Empleando los resultados de los problemas 15.4 y 15.5, la ecuación (*22*) se convierte en

$$x_1 = 0.8546x_2 + 1.5063x_3$$

Dado que $\bar{X}_1 = \dfrac{\sum X_1}{N} = \dfrac{753}{12} = 62.750$ $\bar{X}_2 = \dfrac{\sum X_2}{N} = 53.583$ y $\bar{X}_3 = 8.833$

(de acuerdo con la tabla 15.2 del problema 15.3), la ecuación buscada puede expresarse como

$$X_1 - 62.750 = 0.8546(X_2 - 53.583) + 1.506(X_3 - 8.833)$$

que coincide con el resultado del problema 15.3*a*).

15.7 Dados los datos del problema 15.3, determinar: *a*) el promedio de incremento en el peso por pulgada de incremento en la altura de niños de una misma edad, y *b*) el promedio de incremento en el peso por año en niños de una misma estatura.

SOLUCIÓN

De acuerdo con la ecuación de regresión obtenida en el problema 15.3*a*) o 15.6, se encuentra que la respuesta para *a*) es 0.8546, o bien 0.9 lb, y la respuesta para *b*) es 1.5063, o bien 1.5 lb, aproximadamente.

15.8 Mostrar que las ecuaciones (*3*) y (*4*) de este capítulo se obtienen de las ecuaciones (*1*) y (*2*).

SOLUCIÓN

De acuerdo con la primera de las ecuaciones (*2*), dividiendo ambos lados entre N se tiene

$$\bar{X}_1 = b_{1.23} + b_{12.3}\bar{X}_2 + b_{13.2}\bar{X}_3 \tag{23}$$

Restando la ecuación (*23*) de la ecuación (*1*) se obtiene

$$X_1 - \bar{X}_1 = b_{12.3}(X_2 - \bar{X}_2) + b_{13.2}(X_3 - \bar{X}_3)$$

o bien

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \tag{24}$$

que es la ecuación (*3*).

Sean $X_1 = x_1 + \bar{X}_1$, $X_2 = x_2 + \bar{X}_2$ y $X_3 = x_3 + \bar{X}_3$ en la segunda y tercera ecuación de las ecuaciones (*2*). Entonces, después de algunas simplificaciones algebraicas y empleando los resultados $\sum x_1 = \sum x_2 = \sum x_3 = 0$, estas ecuaciones se convierten en

$$\sum x_1x_2 = b_{12.3}\sum x_2^2 \quad + b_{13.2}\sum x_2x_3 + N\bar{X}_2[b_{1.23} + b_{12.3}\bar{X}_2 + b_{13.2}\bar{X}_3 - \bar{X}_1] \tag{25}$$

$$\sum x_1x_3 = b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2 \quad + N\bar{X}_3[b_{1.23} + b_{12.3}\bar{X}_2 + b_{13.2}\bar{X}_3 - \bar{X}_1] \tag{26}$$

las cuales se reducen a las ecuaciones (*4*) debido a que las cantidades que se encuentran entre corchetes en el lado derecho de las ecuaciones (*25*) y (*26*) son cero de acuerdo con la ecuación (*1*).

15.9 Deducir la ecuación (*5*) que se repite a continuación:

$$\frac{x_1}{s_1} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\frac{x_2}{s_2} + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\frac{x_3}{s_3} \tag{5}$$

SOLUCIÓN

De acuerdo con las ecuaciones (*25*) y (*26*)

$$\begin{aligned} b_{12.3}\sum x_2^2 \quad + b_{13.2}\sum x_2x_3 &= \sum x_1x_2 \\ b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2 \quad &= \sum x_1x_3 \end{aligned} \tag{27}$$

Como

$$s_2^2 = \frac{\sum x_2^2}{N} \qquad \text{y} \qquad s_3^2 = \frac{\sum x_3^2}{N}$$

$\sum x_2^2 = Ns_2^2$ y $\sum x_3^2 = Ns_3^2$. Dado que

$$r_{23} = \frac{\sum x_2x_3}{\sqrt{(\sum x_2^2)(\sum x_3^2)}} = \frac{\sum x_2x_3}{Ns_2s_3}$$

$\sum x_2x_3 = Ns_2s_3r_{23}$. De igual manera, $\sum x_1x_2 = Ns_1s_2r_{12}$ y $\sum x_1x_3 = Ns_1s_3r_{13}$.

Sustituyendo en la ecuación (*27*) y simplificando, se encuentra

$$\begin{aligned} b_{12.3}s_2 \quad &+ b_{13.2}s_3r_{23} = s_1r_{12} \\ b_{12.3}s_2r_{23} + b_{13.2}s_3 \quad &= s_1r_{13} \end{aligned} \tag{28}$$

Resolviendo las ecuaciones simultáneas (*28*), se tiene

$$b_{12.3} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_2}\right) \qquad \text{y} \qquad b_{13.2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_3}\right)$$

Sustituyendo estos valores en la ecuación $x_1 = b_{12.3}x_2 + b_{13.2}x_3$ [ecuación (*24*)] y dividiendo entre s_1 se llega al resultado buscado.

ERROR ESTÁNDAR DE ESTIMACIÓN

15.10 Dados los datos del problema 15.3, calcular el error estándar de estimación de X_1 sobre X_2 y X_3.

SOLUCIÓN

De acuerdo con la tabla 15.3 del problema 15.3, se tiene

$$s_{1.23} = \sqrt{\frac{\sum(X_1 - X_{1,\text{est}})^2}{N}}$$

$$= \sqrt{\frac{(64 - 64.414)^2 + (71 - 69.136)^2 + \cdots + (68 - 65.920)^2}{12}} = 4.6447 \qquad \text{o bien} \qquad 4.6 \text{ lb}$$

El error estándar de estimación poblacional se estima mediante $\hat{s}_{1.23} = \sqrt{N/(N-3)}s_{1.23} = 5.3$ lb en este caso.

15.11 Para obtener el resultado del problema 15.10, utilizar

$$s_{12.3} = s_1\sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

SOLUCIÓN

De acuerdo con los problemas 15.4*a*) y 15.5 se tiene

$$s_{1.23} = 8.6035\sqrt{\frac{1 - (0.8196)^2 - (0.7698)^2 - (0.7984)^2 + 2(0.8196)(0.7698)(0.7984)}{1 - (0.7984)^2}} = 4.6 \text{ lb}$$

Obsérvese que con el método empleado en este problema se obtiene el error estándar de estimación sin necesidad de usar la ecuación de regresión.

COEFICIENTE DE CORRELACIÓN MÚLTIPLE

15.12 Dados los datos del problema 15.3, calcular el coeficiente de correlación lineal múltiple de X_1 sobre X_2 y X_3. Consultar los resultados de MINITAB dados en la solución del problema 15.3 para determinar el coeficiente de correlación lineal múltiple.

SOLUCIÓN

Primer método

De acuerdo con los resultados de los problemas 15.4*a*) y 15.10 se tiene

$$R_{1.23} = \sqrt{1 - \frac{s_{1.23}^2}{s_1^2}} = \sqrt{1 - \frac{(4.6447)^2}{(8.6035)^2}} = 0.8418$$

Segundo método

De acuerdo con los resultados del problema 15.5 se tiene

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} = \sqrt{\frac{(0.8196)^2 + (0.7698)^2 - 2(0.8196)(0.7698)(0.7984)}{1 - (0.7984)^2}} = 0.8418$$

Obsérvese que el coeficiente de correlación múltiple, $R_{1.23}$, es mayor que cualquiera de los coeficientes r_{12} o r_{13} (ver problema 15.5). Esto siempre es así y en realidad es de esperar, ya que al tomar en cuenta más variables independientes relevantes, se llega a una relación mejor entre las variables.

El fragmento siguiente de los resultados de MINITAB en la solución del problema 15.3, `R-Sq = 70.9%`, da el cuadrado del coeficiente de correlación lineal múltiple. El coeficiente de correlación lineal múltiple es la raíz cuadrada de esta cantidad. Es decir, $R_{1.23} = \sqrt{0.709} = 0.842$.

15.13 Dados los datos del problema 15.3, calcular el coeficiente de determinación múltiple de X_1 sobre X_2 y X_3. Consultar los resultados de MINITAB dados en la solución del problema 15.3 para determinar el coeficiente de determinación múltiple.

SOLUCIÓN

El coeficiente de determinación múltiple de X_1 sobre X_2 y X_3 es

$$R_{1.23}^2 = (0.8418)^2 = 0.7086$$

empleando el problema 15.12. Por lo tanto, cerca de 71% de la variación total en X_1 se explica usando la ecuación de regresión.

El coeficiente de determinación múltiple se lee directamente en los resultados de MINITAB dados en la solución del problema 15.3, y es `R-Sq = 70.9%`.

15.14 Según los datos del problema 15.3, calcular: *a*) $R_{2.13}$ y *b*) $R_{3.12}$ y comparar estos valores con el valor de $R_{1.23}$.

SOLUCIÓN

a) $$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}} = \sqrt{\frac{(0.8196)^2 + (0.7984)^2 - 2(0.8196)(0.7698)(0.7984)}{1 - (0.7698)^3}} = 0.8606$$

b) $$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}} = \sqrt{\frac{(0.7698)^2 + (0.7984)^2 - 2(0.8196)(0.7698)(0.7984)}{1 - (0.8196)^2}} = 0.8234$$

Este problema ilustra el hecho de que, en general, $R_{2.13}$, $R_{3.12}$ y $R_{1.23}$ no son necesariamente iguales, como se puede ver en la comparación con el problema 15.12.

15.15 Si $R_{1.23} = 1$, probar que: *a*) $R_{2.13} = 1$ y *b*) $R_{3.12} = 1$.

SOLUCIÓN

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \tag{29}$$

y

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}} \tag{30}$$

a) Haciendo en la ecuación (*29*), $R_{1.23} = 1$ y elevando al cuadrado ambos lados, $r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23} = 1 - r_{23}^2$. Entonces

$$r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23} = 1 - r_{13}^2 \qquad \text{o bien} \qquad \frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2} = 1$$

Es decir, $R_{2.13}^2 = 1$ y $R_{2.13} = 1$, ya que el coeficiente de correlación múltiple se considera no negativo.

b) $R_{3.12} = 1$ sigue del inciso *a*) intercambiando los subíndices 2 y 3 en la fórmula para $R_{2.13} = 1$.

15.16 Si $R_{1.23} = 0$, ¿implica necesariamente que $R_{2.13} = 0$?

SOLUCIÓN

De acuerdo con la ecuación (*29*), $R_{2.13} = 0$ si y sólo si

$$r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23} = 0 \qquad \text{o bien} \qquad 2r_{12}r_{13}r_{23} = r_{12}^2 + r_{13}^2$$

Entonces, de acuerdo con la ecuación (*30*) se tienen

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - (r_{12}^2 + r_{13}^2)}{1 - r_{13}^2}} = \sqrt{\frac{r_{23}^2 - r_{13}^2}{1 - r_{13}^2}}$$

lo cual no es necesariamente igual a cero.

CORRELACIÓN PARCIAL

15.17 Dados los datos del problema 15.3, calcular los coeficientes de correlación lineal parcial $r_{12.3}$, $r_{13.2}$ y $r_{23.1}$. También determinar estos coeficientes empleando STATISTIX.

SOLUCIÓN

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \qquad r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}} \qquad r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}}$$

Empleando los resultados del problema 15.5, se encuentra que $r_{12.3} = 0.5334$, $r_{13.2} = 0.3346$ y $r_{23.1} = 0.4580$. Se concluye que entre los niños de una misma edad, el coeficiente de correlación entre peso y estatura es 0.53; entre los niños de una misma estatura el coeficiente de correlación entre peso y edad es 0.33. Como estos resultados se basan en una muestra pequeña, de sólo 12 niños, no son tan confiables como si se obtuviesen de una muestra mayor.

Con la secuencia **Statistics** → **Linear models** → **Partial Correlations** se obtiene el cuadro de diálogo de la figura 15-4. Este cuadro se llena como se indica en la figura. Se busca $r_{12.3}$. El resultado es el siguiente.

```
Statistix 8.

Partial Correlations with X1
Controlled for X3

X2              0.5335
```

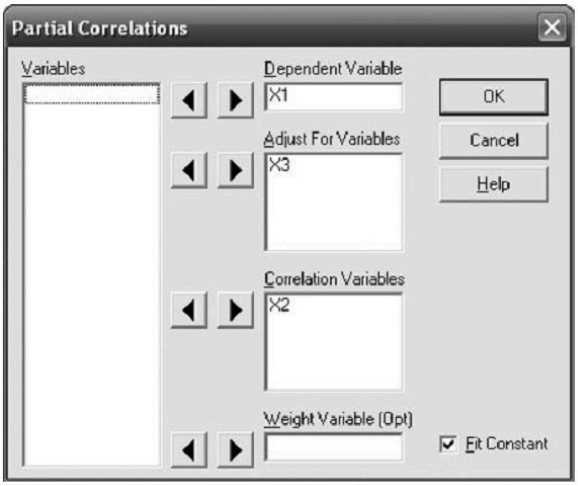


Figura 15-4 STATISTIX, cuadro de diálogo para el problema 15.17.

STATISTIX puede emplearse de manera similar para hallar las otras dos correlaciones parciales buscadas.

15.18 Si $X_1 = b_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$ y $X_3 = b_{3.12} + b_{32.1}X_2 + b_{31.2}X_1$ son las ecuaciones de regresión de X_1 sobre X_2 y X_3, y de X_3 sobre X_2 y X_1, respectivamente, probar que $r^2_{13.2} = b_{13.2}b_{31.2}$.

SOLUCIÓN

La ecuación de regresión de X_1 sobre X_2 y X_3 puede expresarse como [ver ecuación (*5*) de este capítulo]

$$X_1 = \bar{X}_1 = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_3}\right)(X_3 - \bar{X}_3) \qquad (31)$$

La ecuación de regresión de X_3 sobre X_2 y X_1 puede expresarse como [ver ecuación (*10*)]

$$X_3 - \bar{X}_3 = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_3}{s_2}\right)(X_2 - \bar{X}_2) + \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_3}{s_1}\right)(X_1 - \bar{X}_1) \qquad (32)$$

De acuerdo con las ecuaciones (*31*) y (*32*), los coeficientes de X_3 y X_1 son, respectivamente,

$$b_{13.2} = \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\left(\frac{s_1}{s_3}\right) \qquad \text{y} \qquad b_{31.2} = \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right)\left(\frac{s_3}{s_1}\right)$$

Por lo tanto

$$b_{13.2}b_{31.2} = \frac{(r_{13} - r_{12}r_{23})^2}{(1 - r_{23}^2)(1 - r_{12}^2)} = r_{13.2}^2$$

15.19 Si $r_{12.3} = 0$, probar que

$$a) \quad r_{13.2} = r_{13}\sqrt{\frac{1 - r_{23}^2}{1 - r_{12}^2}} \qquad b) \quad r_{23.1} = r_{23}\sqrt{\frac{1 - r_{13}^2}{1 - r_{12}^2}}$$

SOLUCIÓN

Si

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = 0$$

se tiene que $r_{12} = r_{13}r_{23}$.

a) $$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} = \frac{r_{13} - (r_{13}r_{23})r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} = \frac{r_{13}(1-r_{23}^2)}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} = r_{13}\sqrt{\frac{1-r_{23}^2}{1-r_{12}^2}}$$

b) Se intercambian los subíndices 1 y 2 en el resultado del inciso *a*).

CORRELACIÓN MÚLTIPLE Y CORRELACIÓN PARCIAL PARA CUATRO O MÁS VARIABLES

15.20 Un examen de ingreso a la universidad consta de tres partes: matemáticas, español y conocimientos generales. Para determinar si los resultados de este examen sirven para predecir el desempeño en el curso de estadística, se recolectan y se analizan los datos de una muestra de 200 estudiantes. Sea

X_1 = calificación en el curso de estadística $\quad$ X_3 = calificación en el examen de español

X_2 = calificación en el examen de matemáticas $\quad$ X_4 = calificación en el examen de conocimientos generales

Se obtienen los valores siguientes:

$$\bar{X}_1 = 75 \quad s_1 = 10 \quad \bar{X}_2 = 24 \quad s_2 = 5$$
$$\bar{X}_3 = 15 \quad s_3 = 3 \quad \bar{X}_4 = 36 \quad s_4 = 6$$
$$r_{12} = 0.90 \quad r_{13} = 0.75 \quad r_{14} = 0.80 \quad r_{23} = 0.70 \quad r_{24} = 0.70 \quad r_{34} = 0.85$$

Encontrar la ecuación de regresión de mínimos cuadrados de X_1 sobre X_2, X_3 y X_4.

SOLUCIÓN

Generalizando el resultado del problema 15.8, la ecuación de regresión de mínimos cuadrados de X_1 sobre X_2, X_3 y X_4 puede expresarse como

$$x_1 = b_{12.34}x_2 + b_{13.24}x_3 + b_{14.23}x_4 \tag{33}$$

donde $b_{12.34}$, $b_{13.24}$ y $b_{14.23}$ se obtienen a partir de las ecuaciones normales

$$\begin{aligned} \textstyle\sum x_1x_2 &= b_{12.34}\sum x_2^2 + b_{13.24}\sum x_2x_3 + b_{14.23}\sum x_2x_4 \\ \textstyle\sum x_1x_3 &= b_{12.34}\sum x_2x_3 + b_{13.24}\sum x_3^2 + b_{14.23}\sum x_3x_4 \\ \textstyle\sum x_1x_4 &= b_{12.34}\sum x_2x_4 + b_{13.24}\sum x_3x_4 + b_{14.23}\sum x_4^2 \end{aligned} \tag{34}$$

y donde $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$, $x_3 = X_3 - \bar{X}_3$ y $x_4 = X_4 - \bar{X}_4$.

A partir de los datos dados, se encuentra

$$\sum x_2^2 - Ns_2^2 = 5\,000 \quad \sum x_1x_2 = Ns_1s_2r_{12} = 9\,000 \quad \sum x_2x_3 = Ns_1s_3r_{23} = 2\,100$$
$$\sum x_3^2 - Ns_3^2 = 1\,800 \quad \sum x_1x_3 = Ns_1s_3r_{13} = 4\,500 \quad \sum x_2x_4 = Ns_2s_4r_{24} = 4\,200$$
$$\sum x_4^2 - Ns_4^2 = 7\,200 \quad \sum x_1x_4 = Ns_1s_4r_{14} = 9\,600 \quad \sum x_3x_4 = Ns_3s_4r_{34} = 3\,060$$

Sustituyendo estos valores en las ecuaciones (*34*) y resolviendo el sistema de ecuaciones, se obtiene

$$b_{12.34} = 1.3333 \quad b_{13.24} = 0.0000 \quad b_{14.23} = 0.5556 \tag{35}$$

que al sustituirlos en la ecuación (*33*) dan la ecuación de regresión buscada

$$x_1 = 1.3333x_2 + 0.0000x_3 + 0.5556x_4$$

o bien
$$X_1 - 75 = 1.3333(X_2 - 24) + 0.5556(X_4 - 27) \tag{36}$$

o bien
$$X_1 = 22.9999 + 1.3333X_2 + 0.5556X_4$$

La solución exacta de la ecuación (*34*) da $b_{12.34} = \frac{4}{3}$, $b_{13.24} = 0$ y $b_{14.23} = \frac{5}{9}$, de manera que la ecuación de regresión también se puede expresar como

$$X_1 = 23 + \tfrac{4}{3}X_2 + \tfrac{5}{9}X_4 \qquad (37)$$

Es interesante observar que en la ecuación de regresión no aparecen las calificaciones de español, X_3. Esto no significa que los conocimientos de español no sean importantes para el desempeño en estadística, sino que significa que la necesidad del español, en lo que se refiere a la predicción de la calificación en estadística, queda ampliamente reflejada por las calificaciones obtenidas en los otros exámenes.

15.21 Dos estudiantes que aprobaron el examen de admisión del problema 15.20 obtuvieron, respectivamente, las calificaciones siguientes: *a*) 30 en matemáticas, 18 en español y 32 en conocimientos generales y *b*) 18 en matemáticas, 20 en español y 36 en conocimientos generales. ¿Cuál será su calificación en estadística?

SOLUCIÓN

a) Sustituyendo $X_2 = 30$, $X_3 = 18$ y $X_4 = 32$ en la ecuación (*37*), la calificación en estadística será $X_1 = 81$.

b) Procediendo como en el inciso *a*) con $X_2 = 18$, $X_3 = 20$ y $X_4 = 36$, se encuentra $X_1 = 67$.

15.22 Dados los datos del problema 15.20, encontrar los coeficientes de correlación parcial: *a*) $r_{12.34}$, *b*) $r_{13.24}$ y *c*) $r_{14.23}$.

SOLUCIÓN

a) y *b*)
$$r_{12.4} = \frac{r_{12} - r_{14}r_{24}}{\sqrt{(1 - r_{14}^2)(1 - r_{24}^2)}} \qquad r_{13.4} = \frac{r_{13} - r_{14}r_{34}}{\sqrt{(1 - r_{14}^2)(1 - r_{34}^2)}} \qquad r_{23.4} = \frac{r_{23} - r_{24}r_{34}}{\sqrt{(1 - r_{24}^2)(1 - r_{34}^2)}}$$

Sustituyendo con los valores del problema 15.20, se obtiene $r_{12.4} = 0.7935$, $r_{13.4} = 0.2215$ y $r_{23.4} = 0.2791$. Por lo tanto,

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{(1 - r_{13.4}^2)(1 - r_{23.4}^2)}} = 0.7814 \qquad \text{y} \qquad r_{13.24} = \frac{r_{13.4} - r_{12.4}r_{23.4}}{\sqrt{(1 - r_{12.4}^2)(1 - r_{23.4}^2)}} = 0.0000$$

c)
$$r_{14.3} = \frac{r_{14} - r_{13}r_{34}}{\sqrt{(1 - r_{13}^2)(1 - r_{34}^2)}} \qquad r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \qquad r_{24.3} = \frac{r_{24} - r_{23}r_{34}}{\sqrt{(1 - r_{23}^2)(1 - r_{34}^2)}}$$

Sustituyendo con los valores del problema 15.20, se obtiene $r_{14.3} = 0.4664$, $r_{12.3} = 0.7939$ y $r_{24.3} = 0.2791$. Por lo tanto

$$r_{14.23} = \frac{r_{14.3} - r_{12.3}r_{24.3}}{\sqrt{(1 - r_{12.3}^2)(1 - r_{24.3}^2)}} = 0.4193$$

15.23 Interpretar los coeficientes de correlación parcial: *a*) $r_{12.4}$, *b*) $r_{13.4}$, *c*) $r_{12.34}$, *d*) $r_{14.3}$ y *e*) $r_{14.23}$ obtenidos en el problema 15.22.

SOLUCIÓN

a) $r_{12.4} = 0.7935$ representa el coeficiente de correlación (lineal) entre las calificaciones en estadística y las calificaciones en matemáticas de estudiantes con una misma calificación en conocimientos generales. Para obtener este coeficiente no se toman en cuenta las calificaciones en español (así como otros factores tampoco considerados), como resulta evidente por el hecho de que se ha omitido el subíndice 3.

b) $r_{13.4} = 0.2215$ representa el coeficiente de correlación entre las calificaciones en estadística y las calificaciones en español de estudiantes que tienen la misma calificación en conocimientos generales. Aquí no se han considerado las calificaciones en matemáticas.

c) $r_{12.34} = 0.7814$ representa el coeficiente de correlación entre las calificaciones en estadística y las calificaciones en matemáticas de estudiantes con la misma calificación, tanto en español como en conocimientos generales.

d) $r_{14.3} = 0.4664$ representa el coeficiente de correlación entre las calificaciones en estadística y las calificaciones en conocimientos generales de estudiantes con la misma calificación en español.

e) $r_{14.23} = 0.4193$ representa el coeficiente de correlación entre las calificaciones en estadística y las calificaciones en conocimientos generales de estudiantes con la misma calificación tanto en matemáticas como en español.

15.24 *a*) Dados los datos del problema 15.20, mostrar que

$$\frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{(1 - r_{13.4}^2)(1 - r_{23.4}^2)}} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}} \qquad (38)$$

b) Explicar el significado de la igualdad del inciso *a*).

SOLUCIÓN

a) El lado izquierdo de la ecuación (*38*) fue evaluado en el problema 15.22*a*) dando como resultado 0.7814. Para evaluar el lado derecho de la ecuación (*38*), se usan los resultados del problema 15.22*c*); el resultado también es 0.7814. Por lo tanto, en este caso en especial, la igualdad es válida. Mediante manipulaciones algebraicas puede demostrarse que esta igualdad también es válida en general.

b) El lado izquierdo de la ecuación (*38*) es $r_{12.34}$ y el lado derecho es $r_{12.43}$. Como $r_{12.34}$ es la correlación entre las variables X_1 y X_2 cuando X_3 y X_4 permanecen constantes, y $r_{12.43}$ es la correlación entre las variables X_1 y X_2 cuando X_4 y X_3 permanecen constantes, resulta inmediatamente evidente que la igualdad debe ser válida.

15.25 Dados los datos del problema 15.20, encontrar: *a*) el coeficiente de correlación múltiple $R_{1.234}$ y *b*) el error estándar de estimación $s_{1.234}$.

SOLUCIÓN

a) $$1 - R_{1.234}^2 = (1 - r_{12}^2)(1 - r_{13.2}^2)(1 - r_{14.23}^2) \qquad \text{o bien} \qquad R_{1.234} = 0.9310$$

dado que $r_{12} = 0.90$ de acuerdo con el problema 15.20, $r_{14.23} = 0.4193$, de acuerdo con el problema 15.22*c*), y

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}} = \frac{0.75 - (0.90)(0.70)}{\sqrt{[1 - (0.90)^2][1 - (0.70)^2)]}} = 0.3855$$

Otro método

Intercambiando en la primera ecuación los subíndices 2 y 4 se obtiene

$$1 - R_{1.234}^2 = (1 - r_{14}^2)(1 - r_{13.4}^2)(1 - r_{12.34}^2) \qquad \text{o bien} \qquad R_{1.234} = 0.9310$$

donde se han empleado directamente los resultados del problema 15.22*a*).

b) $$R_{1.234} = \sqrt{\frac{1 - s_{1.234}^2}{s_1^2}} \qquad \text{o bien} \qquad s_{1.234} = s_1\sqrt{1 - R_{1.234}^2} = 10\sqrt{1 - (0.9310)^2} = 3.650$$

Comparar con la ecuación (*8*) de este capítulo.

PROBLEMAS SUPLEMENTARIOS

ECUACIONES DE REGRESIÓN CON TRES VARIABLES

15.26 Empleando la notación adecuada con subíndices, escribir las ecuaciones de regresión de: *a*) X_3 sobre X_1 y X_2, y *b*) X_4 sobre X_1, X_2, X_3 y X_5.

15.27 Escribir las ecuaciones normales correspondientes a la ecuación de regresión de: *a*) X_2 sobre X_1 y X_3, y *b*) X_5 sobre X_1, X_2, X_3 y X_4.

15.28 En la tabla 15.4 se presentan los valores de tres variables: X_1, X_2 y X_3.

a) Encontrar la ecuación de regresión de mínimos cuadrados de X_3 sobre X_1 y X_2.

b) Estimar X_3 para $X_1 = 10$ y $X_2 = 6$.

Tabla 15.4

X_1	3	5	6	8	12	14
X_2	16	10	7	4	3	2
X_3	90	72	54	42	30	12

15.29 Un maestro de matemáticas quiere determinar la relación que hay entre las calificaciones del examen final y las calificaciones de dos exámenes parciales durante el semestre. Siendo X_1, X_2 y X_3, respectivamente, las calificaciones del primero y segundo exámenes parciales y del examen final, el profesor calcula los siguientes valores correspondientes a un total de 120 alumnos.

$$\bar{X}_1 = 6.8 \qquad \bar{X}_2 = 7.0 \qquad \bar{X}_3 = 74$$
$$s_1 = 1.0 \qquad s_2 = 0.80 \qquad s_2 = 9.0$$
$$r_{12} = 0.60 \qquad r_{13} = 0.70 \qquad r_{23} = 0.65$$

a) Encontrar la ecuación de regresión de mínimos cuadrados de X_3 sobre X_1 y X_2.

b) Estimar la calificación final de dos estudiantes cuyas calificaciones en los dos exámenes parciales fueron: 1) 9 y 7, y 2) 4 y 8.

15.30 Los datos de la tabla 15.5 dan el precio en miles (X_1), la cantidad de recámaras (X_2) y la cantidad de baños (X_3) de 10 casas. Usar las ecuaciones normales para hallar la ecuación de regresión de mínimos cuadrados de X_1 sobre X_2 y X_3. Usar EXCEL, MINITAB, SAS, SPSS y STATISTIX para encontrar la ecuación de regresión de mínimos cuadrados de X_1 sobre X_2 y X_3. Usar la ecuación de regresión de mínimos cuadrados de X_1 sobre X_2 y X_3 para estimar el precio de una casa que tenga cinco recámaras y cuatro baños.

Tabla 15.5

Precio	Recámaras	Baños
165	3	2
200	3	3
225	4	3
180	2	3
202	4	2
250	4	4
275	3	4
300	5	3
155	2	2
230	4	4

ERROR ESTÁNDAR DE ESTIMACIÓN

15.31 Dados los datos del problema 15.28, encontrar el error estándar de estimación de X_3 sobre X_1 y X_2.

15.32 Dados los datos del problema 15.29, encontrar el error estándar de estimación de: *a*) X_3 sobre X_1 y X_2 y *b*) X_1 sobre X_2 y X_3.

COEFICIENTE DE CORRELACIÓN MÚLTIPLE

15.33 Dados los datos del problema 15.28, calcular el coeficiente de correlación lineal múltiple de X_3 sobre X_1 y X_2.

15.34 Dados los datos del problema 15.29, calcular: *a*) $R_{3.12}$, *b*) $R_{1.23}$ y *c*) $R_{2.13}$.

15.35 *a*) Si $r_{12} = r_{13} = r_{23} = r \neq 1$, mostrar que

$$R_{1.23} = R_{2.31} = R_{3.12} = \frac{r\sqrt{2}}{\sqrt{1+r}}$$

b) Analizar el caso $r = 1$.

15.36 Si $R_{1.23} = 0$, probar que $|r_{23}| \geq |r_{12}|$ y $|r_{23}| \geq |r_{13}|$ e interpretar.

CORRELACIÓN PARCIAL

15.37 Dados los datos del problema 15.28, calcular los coeficientes de correlación lineal parcial $r_{12.3}$, $r_{13.2}$ y $r_{23.1}$. Calcularlos también usando STATISTIX.

15.38 Resolver el problema 15.37 con los datos del problema 15.29.

15.39 Si $r_{12} = r_{13} = r_{23} = r \neq 1$, mostrar que $r_{12.3} = r_{13.2} = r_{23.1} = r/(1+r)$. Analizar el caso $r = 1$.

15.40 Si $r_{12.3} = 1$, mostrar que: *a*) $|r_{13.2}| = 1$, *b*) $|r_{23.1}| = 1$, *c*) $R_{1.23} = 1$ y *d*) $s_{1.23} = 0$.

CORRELACIÓN MÚLTIPLE Y CORRELACIÓN PARCIAL CON CUATRO O MÁS VARIABLES

15.41 Mostrar que la ecuación de regresión de X_4 sobre X_1, X_2 y X_3 puede escribirse como

$$\frac{x_4}{s_4} = a_1\left(\frac{x_1}{s_1}\right) + a_2\left(\frac{x_2}{s_2}\right) + a_3\left(\frac{x_3}{s_3}\right)$$

donde a_1, a_2 y a_3 se determinan resolviendo simultáneamente las ecuaciones

$$a_1 r_{11} + a_2 r_{12} + a_3 r_{13} = r_{14}$$
$$a_1 r_{21} + a_2 r_{22} + a_3 r_{23} = r_{24}$$
$$a_1 r_{31} + a_2 r_{32} + a_3 r_{33} = r_{34}$$

y donde $x_j = X_j - \bar{X}_j$, $r_{jj} = 1$ y $j = 1, 2, 3$ y 4. Generalizar al caso con más de cuatro variables.

15.42 Dados $\bar{X}_1 = 20$, $\bar{X}_2 = 36$, $\bar{X}_3 = 12$, $\bar{X}_4 = 80$, $s_1 = 1.0$, $s_2 = 2.0$, $s_3 = 1.5$, $s_4 = 6.0$, $r_{12} = -0.20$, $r_{13} = 0.40$, $r_{23} = 0.50$, $r_{14} = 0.40$, $r_{24} = 0.30$ y $r_{34} = -0.10$, *a*) encontrar la ecuación de regresión de X_4 sobre X_1, X_2 y X_3, y *b*) estimar X_4 para $X_1 = 15$, $X_2 = 40$ y $X_3 = 14$.

15.43 Dados los datos del problema 15.42, encontrar: *a*) $r_{41.23}$, *b*) $r_{42.13}$ y *c*) $r_{43.12}$.

15.44 Dados los datos del problema 15.42, encontrar: *a*) $R_{4.123}$ y *b*) $s_{4.123}$.

15.45 Los gastos médicos anuales de quince hombres adultos se correlacionan con otros factores de salud. En un estudio se consideran gastos médicos anuales, Y, así como la información sobre las siguientes variables independientes,

$$X_1 = \begin{cases} 0, & \text{si es no fumador} \\ 1, & \text{si es fumador} \end{cases} \qquad X_2 = \text{cantidad de dinero gastado semanalmente en alcohol,}$$

$$X_3 = \text{horas semanales de ejercicio,}$$

$$X_4 = \begin{cases} 0, \text{ poco informado sobre la alimentación} \\ 1, \text{ informado medianamente sobre la alimentación} \\ 2, \text{ altamente informado sobre la alimentación} \end{cases}$$

$$X_5 = \text{peso} \qquad X_6 = \text{edad}$$

La notación empleada en este problema se encuentra en muchos libros de estadística. Y se emplea como variable dependiente y X, con subíndices, como variables independientes. Empleando los datos de la tabla 15.6, encontrar, resolviendo las ecuaciones normales, la ecuación de regresión de Y sobre X_1 a X_6 y comparar esta solución con las soluciones dadas por EXCEL, MINITAB, SAS, SPSS y STATISTIX.

Tabla 15.6

Gastos médicos	Fumador	Alcohol	Ejercicio	Alimentación	Peso	Edad
2 100	0	20	5	1	185	50
2 378	1	25	0	1	200	42
1 657	0	10	10	2	175	37
2 584	1	20	5	2	225	54
2 658	1	25	0	1	220	32
1 842	0	0	10	1	165	34
2 786	1	25	5	0	225	30
2 178	0	10	10	1	180	41
3 198	1	30	0	1	225	31
1 782	0	5	10	0	180	45
2 399	0	25	12	2	225	45
2 423	0	15	15	0	220	33
3 700	1	25	0	1	275	43
2 892	1	30	5	1	230	42
2 350	1	30	10	1	245	40

ANÁLISIS DE VARIANZA

16

OBJETIVO DEL ANÁLISIS DE VARIANZA

En el capítulo 8 se usó la teoría del muestreo para probar la importancia de la diferencia entre dos medias muestrales y se supuso que las dos poblaciones de las que provenían las muestras tenían la misma varianza. Hay ocasiones que se necesita probar la importancia de la diferencia entre tres o más medias muestrales o, lo que es equivalente, probar la hipótesis nula de que todas estas medias muestrales son iguales.

EJEMPLO 1 Supóngase que en un experimento agrícola se emplean cuatro diferentes tratamientos químicos para el suelo, y se obtienen, respectivamente, con los siguientes rendimientos medios de trigo: 28, 22, 18 y 24 bushels por acre. ¿Existe diferencia significativa entre estas medias o la dispersión observada se debe sólo a la casualidad?

Problemas como éste se resuelven empleando una técnica desarrollada por Fischer y que se denomina *análisis de varianza*. En esta técnica se usa la distribución F, ya vista en el capítulo 11.

CLASIFICACIÓN EN UN SENTIDO O EXPERIMENTOS CON UN FACTOR

En un *experimento de un factor*, las mediciones (u observaciones) se hacen de a grupos independientes de muestras, y b es la cantidad de mediciones en cada muestra. Se habla de *a tratamientos*, cada uno con *b repeticiones* o *b réplicas*. En el ejemplo 1, $a = 4$.

Los resultados de un experimento de un factor se acostumbra presentarlos en una tabla con a renglones y b columnas, como la tabla 16.1. Aquí, X_{jk} denota la medición del renglón j y columna k, donde $j = 1, 2, \ldots, a$ y donde $k = 1, 2, \ldots, b$. Por ejemplo, X_{35} significa la quinta medición del tercer tratamiento.

Tabla 16.1

Tratamiento 1	$X_{11}, X_{12}, \ldots, X_{1b}$	$\bar{X}_{1.}$
Tratamiento 2	$X_{21}, X_{22}, \ldots, X_{2b}$	$\bar{X}_{2.}$
⋮	⋮	⋮
Tratamiento a	$X_{a1}, X_{a2}, \ldots, X_{ab}$	$\bar{X}_{a.}$

La media de las mediciones en el renglón j se denota $\bar{X}_{j.}$. Se tiene

$$\bar{X}_{j.} = \frac{1}{b}\sum_{k=1}^{b} X_{jk} \qquad j = 1, 2, \ldots, a \tag{1}$$

El punto que aparece en $\bar{X}_{j.}$ sirve para indicar que se suma sobre el índice k. A los valores $X_{j.}$ se les llama *medias de grupo, medias de tratamiento* o *medias de renglón*. La *gran media* o *media general* es la media de todas las mediciones de todos los grupos y se denota $\bar{X}$:

$$\bar{X} = \frac{1}{ab}\sum_{j=1}^{a}\sum_{k=1}^{b} X_{jk} \tag{2}$$

VARIACIÓN TOTAL, VARIACIÓN DENTRO DE TRATAMIENTOS Y VARIACIÓN ENTRE TRATAMIENTOS

La *variación total*, que se denota V, se define como la suma de los cuadrados de las desviaciones de cada medición respecto a la gran media $\bar{X}$

$$\text{Variación total} = V = \sum_{j,k}(X_{jk} - \bar{X})^2 \tag{3}$$

Expresando esta identidad como

$$X_{jk} - \bar{X} = (X_{jk} - \bar{X}_{j.}) + (\bar{X}_{j.} - \bar{X}) \tag{4}$$

y después elevando al cuadrado y sumando sobre j y k, se tiene (ver problema 16.1)

$$\sum_{j,k}(X_{jk} - \bar{X})^2 = \sum_{j,k}(X_{jk} - \bar{X}_{j.})^2 + \sum_{j,k}(\bar{X}_{j.} - \bar{X})^2 \tag{5}$$

o

$$\sum_{j,k}(X_{jk} - \bar{X})^2 = \sum_{j,k}(X_{jk} - \bar{X}_{j.})^2 + b\sum_{j}(\bar{X}_{j.} - \bar{X})^2 \tag{6}$$

La primera suma que aparece en el lado derecho de las ecuaciones (*5*) y (*6*) es la *variación dentro de los tratamientos* (ya que se trata de los cuadrados de las desviaciones de las X_{jk} respecto a las medias de los tratamientos $\bar{X}_{j.}$) y se denota V_W. Por lo tanto,

$$V_W = \sum_{j,k}(X_{jk} - \bar{X}_{j.})^2 \tag{7}$$

La segunda suma que aparece en el lado derecho de las ecuaciones (*5*) y (*6*) es la *variación entre los tratamientos* (ya que se trata de los cuadrados de las desviaciones de las medias de los tratamientos $\bar{X}_{j.}$ respecto a la gran media $\bar{X}$) y se denota V_B. Por lo tanto,

$$V_B = \sum_{j,k}(\bar{X}_{j.} - \bar{X})^2 = b\sum_{j}(\bar{X}_{j} - \bar{X})^2 \tag{8}$$

Por lo tanto, las ecuaciones (*5*) y (*6*) se pueden expresar como

$$V = V_W + V_B \tag{9}$$

MÉTODOS ABREVIADOS PARA OBTENER LAS VARIACIONES

Para simplificar el cálculo de las variaciones anteriores se emplean las fórmulas siguientes:

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} \tag{10}$$

$$V_B = \frac{1}{b}\sum_{j} T_{j.}^2 - \frac{T^2}{ab} \tag{11}$$

$$V_W = V - V_B \tag{12}$$

donde T es la suma de todos los valores X_{jk} y donde $T_{j.}$ es la suma de todos los valores del tratamiento j-ésimo:

$$T = \sum_{j,k} X_{jk} \qquad T_{j.} = \sum_{k} X_{jk} \tag{13}$$

En la práctica, conviene sustraer, de cada dato de la tabla, un valor fijo con objeto de simplificar los cálculos; esto no afecta el resultado final.

MODELO MATEMÁTICO PARA EL ANÁLISIS DE VARIANZA

Cada renglón de la tabla 16.1 se considera como una muestra aleatoria de tamaño b tomada de la población de ese determinado tratamiento. Las X_{jk} difieren de la media poblacional μ_j correspondiente al tratamiento j en un *error aleatorio* que se denota ε_{jk}; por lo tanto,

$$X_{jk} = \mu_j + \varepsilon_{jk} \tag{14}$$

Se supone que estos errores están distribuidos de manera normal con media 0 y varianza σ^2. Si μ es la media de la población de todos los tratamientos y si se denota $\alpha_j = \mu_j - \mu$, entonces $\mu_j = \mu + \alpha_j$, y la ecuación (*14*) se convierte en

$$X_{jk} = \mu + \alpha_j + \varepsilon_{jk} \tag{15}$$

donde $\sum_j \alpha_j = 0$ (ver problema 16.18). De acuerdo con la ecuación (*15*) y con la suposición de que las ε_{jk} están distribuidas de manera normal con media 0 y varianza σ^2, se concluye que las X_{jk} se pueden considerar como variables aleatorias distribuidas en forma normal, con media μ y varianza σ^2.

La hipótesis nula de que todas las medias de los tratamientos son iguales está dada por ($H_0 : \alpha_j = 0; j = 1, 2, \ldots, a$) o, lo que es equivalente, por ($H_0 : \mu_j = \mu; j = 1, 2, \ldots, a$). Si H_0 es verdadera, todas las poblaciones de los tratamientos tendrán la misma distribución normal (es decir, con la misma media y varianza). En estos casos, sólo hay un tratamiento poblacional (es decir, todos los tratamientos son estadísticamente idénticos); en otras palabras, no hay diferencia significativa entre los tratamientos.

VALORES ESPERADOS DE LAS VARIACIONES

Como se puede demostrar (ver problema 16.19), los valores esperados de V_W, V_B y V están dados por

$$E(V_W) = a(b-1)\sigma^2 \tag{16}$$

$$E(V_B) = (a-1)\sigma^2 + b\sum_j \alpha_j^2 \tag{17}$$

$$E(V) = (ab-1)\sigma^2 + b\sum_j \alpha_j^2 \tag{18}$$

De acuerdo con la ecuación (*16*) se tiene

$$E\left[\frac{V_W}{a(b-1)}\right] = \sigma^2 \tag{19}$$

de manera que

$$\hat{S}_W^2 = \frac{V_W}{a(b-1)} \tag{20}$$

siempre es la mejor estimación (insesgada) de σ^2, sin importar si H_0 es o no verdadera. Por otro lado, de acuerdo con las ecuaciones (*17*) y (*18*) se ve que sólo si H_0 es verdadera (es decir, $\alpha_j = 0$) se tendrá

$$E\left(\frac{V_B}{a-1}\right) = \sigma^2 \qquad \text{y} \qquad E\left(\frac{V}{ab-1}\right) = \sigma^2 \tag{21}$$

de manera que sólo en ese caso

$$\hat{S}_B^2 = \frac{V_B}{a-1} \qquad \text{y} \qquad \hat{S}^2 = \frac{V}{ab-1} \tag{22}$$

proporcionan una estimación insesgada de σ^2. Pero si H_0 no es verdadera, entonces de acuerdo con la ecuación (*17*) se tiene

$$E(\hat{S}_B^2) = \sigma^2 + \frac{b}{a-1}\sum_j \alpha_j^2 \tag{23}$$

DISTRIBUCIONES DE LAS VARIACIONES

Empleando la propiedad aditiva de ji cuadrada se pueden probar los siguientes teoremas fundamentales que se refieren a las distribuciones de las variaciones V_W, V_B y V:

Teorema 1: V_W/σ^2 tienen una distribución ji cuadrada con $a(b-1)$ grados de libertad.

Teorema 2: Bajo la hipótesis nula H_0, V_B/σ^2 y V/σ^2 tienen distribuciones ji cuadrada con $a-1$ y $ab-1$ grados de libertad, respectivamente.

Es importante subrayar que el teorema 1 es válido, ya sea que H_0 sea o no verdadera, mientras que el teorema 2 sólo es válido bajo la suposición de que H_0 es verdadera.

PRUEBA *F* PARA LA HIPÓTESIS NULA DE MEDIAS IGUALES

Si la hipótesis nula H_0 no es verdadera (es decir, si las medias de los tratamientos no son iguales), como se ve de acuerdo con la ecuación (*23*), se esperará que $\hat{S}_B^2$ sea mayor que σ^2, y que este efecto se haga más pronunciado a medida que la discrepancia entre las medias aumente. Por otro lado, de acuerdo con las ecuaciones (*19*) y (*20*) puede esperarse que $\hat{S}_W^2$ sea igual a σ^2 sin importar si las medias son o no iguales. Se tiene, entonces, que un buen estadístico para probar la hipótesis H_0 es el proporcionado por $\hat{S}_B^2/\hat{S}_W^2$. Si este estadístico es significativamente grande, se puede concluir que entre las medias de los tratamientos hay una diferencia significativa y, por lo tanto, se puede rechazar H_0; si no es así, puede aceptarse H_0 o posponer la decisión hasta hacer más análisis.

Para usar el estadístico $\hat{S}_B^2/\hat{S}_W^2$ es preciso conocer su distribución muestral. Este conocimiento lo proporciona el teorema 3.

Teorema 3: El estadístico $F = \hat{S}_B^2/\hat{S}_W^2$ tiene distribución F con $a-1$ y $a(b-1)$ grados de libertad.

El teorema 3 permite probar la hipótesis nula a determinado nivel de significancia, empleando la distribución F (estudiada en el capítulo 11) mediante una prueba de una cola.

TABLAS PARA EL ANÁLISIS DE VARIANZA

En la tabla 16.2, llamada *tabla para el análisis de varianza*, se resumen los cálculos necesarios para la prueba anterior. En la práctica se calculan V y V_B empleando ya sea el método largo [ecuaciones (*3*) y (*8*)] o el método corto [ecuaciones (*10*) y (*11*)] y calculando después $V_W = V - V_B$. Debe notarse que el número de grados de libertad para la variación total (es decir, $ab-1$) es igual a la suma de los grados de libertad para la variación entre los tratamientos más los grados de libertad para la variación dentro de los tratamientos.

Tabla 16.2

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, $V_B = b\sum_j (\bar{X}_{j.} - \bar{X})^2$	$a-1$	$\hat{S}_B^2 = \dfrac{V_B}{a-1}$	$\dfrac{\hat{S}_B^2}{\hat{S}_W^2}$
Dentro de tratamientos, $V_W = V - V_B$	$a(b-1)$	$\hat{S}_W^2 = \dfrac{V_W}{a(b-1)}$	con $a-1$ y $a(b-1)$ grados de libertad
Total $V = V_B + V_W = \sum_{j,k} (X_{jk} - \bar{X})^2$	$ab-1$		

MODIFICACIONES PARA NÚMEROS DISTINTOS DE OBSERVACIONES

En caso de que los tratamientos $1, \ldots, a$ tengan números distintos de observaciones —iguales a $N_1, \ldots, N_a$, respectivamente— los resultados anteriores pueden modificarse fácilmente. Así se obtiene

$$V = \sum_{j,k} (X_{jk} - \bar{X})^2 = \sum_{j,k} X_{jk}^2 - \frac{T^2}{N} \tag{24}$$

$$V_B = \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 = \sum_{j} N_j (\bar{X}_{j.} - \bar{X})^2 = \sum_{j} \frac{T_{j.}^2}{N_j} - \frac{T^2}{N} \tag{25}$$

$$V_W = V - V_B \tag{26}$$

donde $\sum_{j,k}$ denota la sumatoria, primero sobre k desde 1 hasta N_j y después la sumatoria sobre j desde 1 hasta a. En este caso, la tabla para el análisis de varianza es la tabla 16.3.

Tabla 16.3

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, $V_B = \sum_j N_j (\bar{X}_{j.} - \bar{X})^2$	$a-1$	$\hat{S}_B^2 = \dfrac{V_B}{a-1}$	$\dfrac{\hat{S}_B^2}{\hat{S}_W^2}$
Dentro de tratamientos, $V_W = V - V_B$	$N-a$	$\hat{S}_W^2 = \dfrac{V_W}{N-a}$	con $a-1$ y $N-a$ grados de libertad
Total, $V = V_B + V_W = \sum_{j,k} (X_{jk} - \bar{X})^2$	$N-1$		

CLASIFICACIÓN EN DOS SENTIDOS O EXPERIMENTOS CON DOS FACTORES

Las ideas del análisis de varianza para clasificaciones en un sentido o experimentos con un factor pueden generalizarse. En el ejemplo 2 se ilustra el procedimiento para *clasificaciones en dos sentidos* o *experimentos con dos factores*.

EJEMPLO 2 Supóngase que un experimento agrícola consiste en examinar los rendimientos por acre de cuatro variedades de trigo, cultivando cada variedad en cinco tipos de parcelas. Por lo tanto, se necesitarán (4)(5) = 20 parcelas. En tales casos conviene reunir las parcelas en *bloques*, por ejemplo, bloques de cuatro parcelas, y cultivar una variedad diferente de trigo en cada parcela del bloque. Así, en este ejemplo se necesitarán 5 bloques.

En este caso se tienen dos clasificaciones, o dos factores, ya que las diferencias en el rendimiento por acre pueden deberse a: 1) el tipo de trigo cultivado, o 2) al bloque de que se trate (que pueden presentar diferencias en la fertilidad del suelo, etcétera).

Por analogía, con el experimento agrícola del ejemplo 2 se acostumbra referirse a los dos factores de un experimento como *tratamientos* y *bloques*, aunque por supuesto puede referirse a ellos simplemente como factor 1 y factor 2.

NOTACIÓN PARA EXPERIMENTOS CON DOS FACTORES

Cuando se tienen a tratamientos y b bloques, se construye una tabla como la 16.4, donde se supone que para cada tratamiento y para cada bloque hay un valor experimental (por ejemplo, el rendimiento por acre). X_{jk} denota el tratamiento j y el bloque k. La media de las entradas en el renglón j se denota $\bar{X}_{j.}$, donde $j = 1, \ldots, a$, y la media de las entradas en la columna k se denota $\bar{X}_{.k}$, donde $k = 1, \ldots, b$. La media general, o gran media, se denota $\bar{X}$. En símbolos,

$$\bar{X}_{j.} = \frac{1}{b}\sum_{k=1}^{b} X_{jk} \qquad \bar{X}_{.k} = \frac{1}{a}\sum_{j=1}^{a} X_{jk} \qquad \bar{X} = \frac{1}{ab}\sum_{j,k} X_{jk} \tag{27}$$

Tabla 16.4

	Bloque				
	1	2	...	b	
Tratamiento 1	X_{11}	X_{12}	...	X_{1b}	$\bar{X}_{1.}$
Tratamiento 2	X_{21}	X_{22}	...	X_{2b}	$\bar{X}_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮
Tratamiento a	X_{a1}	X_{a2}	...	X_{ab}	$\bar{X}_{a.}$
	$\bar{X}_{.1}$	$\bar{X}_{.2}$		$\bar{X}_{.b}$	

VARIACIONES EN LOS EXPERIMENTOS CON DOS FACTORES

Como en el caso de los experimentos con un factor, se definen las variaciones en los experimentos con dos factores. Primero, como en la ecuación (*3*), se define la *variación total*, que es

$$V = \sum_{j,k}(X_{jk} - \bar{X})^2 \tag{28}$$

Expresando la identidad

$$X_{jk} - \bar{X} = (X_{jk} - \bar{X}_{j.} - \bar{X}_{.k} + \bar{X}) + (\bar{X}_{j.} - \bar{X}) + (\bar{X}_{.k} - \bar{X}) \tag{29}$$

elevando al cuadrado y sumando después sobre j y k se puede mostrar que

$$V = V_E + V_R + V_C \tag{30}$$

donde V_E = variación debida al error o a la casualidad $= \sum_{j,k}(X_{jk} - \bar{X}_{j.} - \bar{X}_{.k} + \bar{X})^2$

V_R = variación entre renglones (tratamientos) $= b\sum_{j=1}^{a}(\bar{X}_{j.} - \bar{X})^2$

V_C = variación entre columnas (bloques) $= a\sum_{k=1}^{b}(\bar{X}_{.k} - \bar{X})^2$

La variación debida al error o a la casualidad se conoce también como *variación residual* o *variación aleatoria*.

Las fórmulas siguientes, análogas a las ecuaciones (*10*), (*11*) y (*12*), son las fórmulas de cálculo abreviadas.

$$V = \sum_{jk} X_{jk}^2 - \frac{T^2}{ab} \tag{31}$$

$$V_R = \frac{1}{b}\sum_{j=1}^{a} T_{j.}^2 - \frac{T^2}{ab} \tag{32}$$

$$V_C = \frac{1}{a}\sum_{k=1}^{b} T_{.k}^2 - \frac{T^2}{ab} \tag{33}$$

$$V_E = V - V_R - V_C \tag{34}$$

donde T_j es el total (la suma) de las entradas en el renglón j-ésimo, $T_{.k}$ es el total (la suma) de las entradas en la columna k, y T es el total (la suma) de todas las entradas.

ANÁLISIS DE VARIANZA PARA EXPERIMENTOS CON DOS FACTORES

La generalización del modelo matemático para experimentos con un factor, dado por la ecuación (*15*), lleva a suponer que para experimentos con dos factores

$$X_{jk} = \mu + \alpha_j + \beta_k + \varepsilon_{jk} \tag{35}$$

donde $\sum \alpha_j = 0$ y $\sum \beta_k = 0$. Aquí μ es la gran media de la población, α_j es la parte de X_{jk} atribuida a los diferentes tratamientos (también llamada *efectos del tratamiento*), β_k es la parte de X_{jk} atribuida a los diferentes bloques (también llamada *efectos de los bloques*) y ε_{jk} es la parte de X_{jk} atribuida a la casualidad o al error. Como antes, se supone que las ε_{jk} están distribuidas en forma normal con media 0 y varianza σ^2, de manera que las X_{jk} también están distribuidas en forma normal con media μ y varianza σ^2.

Correspondiendo con los resultados (*16*), (*17*) y (*18*) puede probarse que las esperanzas de las variaciones están dadas por

$$E(V_E) = (a-1)(b-1)\sigma^2 \tag{36}$$

$$E(V_R) = (a-1)\sigma^2 + b\sum_j \alpha_j^2 \tag{37}$$

$$E(V_C) = (b-1)\sigma^2 + a\sum_k \beta_k^2 \tag{38}$$

$$E(V) = (ab-1)\sigma^2 + b\sum_j \alpha_j^2 + a\sum_k \beta_k^2 \tag{39}$$

Las hipótesis nulas que se quieren probar son dos:

$H_0^{(1)}$: todas las medias de los tratamientos (renglones) son iguales; es decir, $\alpha_j = 0$ y $j = 1, \ldots, a$.

$H_0^{(2)}$: todas las medias de los bloques (columnas) son iguales; es decir, $\beta_k = 0$ y $k = 1, \ldots, b$.

De acuerdo con la ecuación (*36*) se ve que, ya sea que $H_0^{(1)}$ y $H_0^{(2)}$, sean o no verdaderas, una estimación insesgada de σ^2 es la dada por

$$\hat{S}_E^2 = \frac{V_E}{(a-1)(b-1)} \qquad \text{esto es,} \qquad E(\hat{S}_E^2) = \sigma^2 \tag{40}$$

Además, si las hipótesis $H_0^{(1)}$ y $H_0^{(2)}$ son verdaderas, entonces

$$\hat{S}_R^2 = \frac{V_R}{a-1} \qquad \hat{S}_C^2 = \frac{V_C}{b-1} \qquad \hat{S}^2 = \frac{V}{ab-1} \tag{41}$$

serán estimaciones insesgadas de σ^2. Sin embargo, si $H_0^{(1)}$ y $H_0^{(2)}$ no son verdaderas, de acuerdo con las ecuaciones (*37*) y (*38*) se tiene, respectivamente

$$E(\hat{S}_R^2) = \sigma^2 + \frac{b}{a-1}\sum_j \alpha_j^2 \tag{42}$$

$$E(\hat{S}_C^2) = \sigma^2 + \frac{a}{b-1}\sum_k \beta_k^2 \tag{43}$$

Los teoremas siguientes son similares a los teoremas 1 y 2:

Teorema 4: V_E/σ^2 es una distribución ji cuadrada con $(a-1)(b-1)$ grados de libertad, independientemente de $H_0^{(1)}$ o bien $H_0^{(2)}$.

Teorema 5: Si la hipótesis $H_0^{(1)}$ es verdadera, V_R/σ^2 tiene una distribución ji cuadrada con $a-1$ grados de libertad. Si la hipótesis $H_0^{(2)}$ es verdadera, V_C/σ^2 tiene una distribución ji cuadrada con $b-1$ grados de libertad. Si las dos hipótesis $H_0^{(1)}$ y $H_0^{(2)}$ son verdaderas, V/σ^2 es una distribución ji cuadrada con $ab-1$ grados de libertad.

Para probar la hipótesis $H_0^{(1)}$, es natural considerar el estadístico $\hat{S}_R^2/\hat{S}_E^2$, ya que, como se ve, de acuerdo con la ecuación (*42*), $\hat{S}_R^2$ se espera que difiera significativamente de σ^2 si las medias de los renglones (tratamientos) son significativamente diferentes. De manera similar, para probar la hipótesis $H_0^{(2)}$, se emplea el estadístico $\hat{S}_C^2/\hat{S}_E^2$. En el teorema 6 se dan las distribuciones de $\hat{S}_R^2/\hat{S}_E^2$ y de $\hat{S}_C^2/\hat{S}_E^2$; este teorema es análogo al teorema 3.

Teorema 6: Si la hipótesis $H_0^{(1)}$ es verdadera, el estadístico $\hat{S}_R^2/\hat{S}_E^2$ tiene una distribución F con $a-1$ y $(a-1)(b-1)$ grados de libertad. Si la hipótesis $H_0^{(2)}$ es verdadera, el estadístico $\hat{S}_C^2/\hat{S}_E^2$ tiene la distribución F con $b-1$ y $(a-1)(b-1)$ grados de libertad.

El teorema 6 permite aceptar o rechazar $H_0^{(1)}$ y $H_0^{(2)}$ a un nivel de significancia determinado. Para mayor claridad y facilidad, como en el caso de un factor, para el análisis de varianza se suele construir una tabla como la 16.5.

EXPERIMENTOS CON DOS FACTORES CON REPLICACIÓN

En la tabla 16.4, para cada tratamiento y para cada bloque hay únicamente una entrada. Más información acerca de los factores puede obtenerse repitiendo el experimento, proceso que se llama *replicación*. En esos casos habrá más de una entrada para cada tratamiento y para cada bloque. Se supondrá que en cada posición hay c entradas; en el caso en que los números de replicaciones no sean iguales, se hacen las modificaciones apropiadas.

Debido a la replicación se necesita un modelo adecuado que sustituya al modelo dado por la ecuación (*35*). Se usa el modelo siguiente:

$$X_{jkl} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{jkl} \tag{44}$$

donde los subíndices j, k y l de X_{jkl} corresponden, respectivamente, al j-ésimo renglón (o tratamientos), a la k-ésima columna (o bloque) y a la l-ésima repetición (o replicación). En la ecuación (*44*) μ, α_j y β_k están definidos como antes;

Tabla 16.5

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, $V_R = b\sum_j (\bar{X}_{j.} - \bar{X})^2$	$a-1$	$\hat{S}_R^2 = \dfrac{V_R}{a-1}$	$\hat{S}_R^2/\hat{S}_E^2$ con $a-1$ y $(a-1)(b-1)$ grados de libertad
Entre bloques, $V_C = a\sum_k (\bar{X}_{.k} - \bar{X})^2$	$b-1$	$\hat{S}_C^2 = \dfrac{V_C}{b-1}$	$\hat{S}_C^2/\hat{S}_E^2$ con $b-1$ y $(a-1)(b-1)$ grados de libertad
Residual o aleatoria, $V_E = V - V_R - V_C$	$(a-1)(b-1)$	$\hat{S}_E^2 = \dfrac{V_E}{(a-1)(b-1)}$	
Total, $V = V_R + V_C + V_E = \sum_{j,k} (X_{jk} - \bar{X})^2$	$ab-1$		

ε_{jkl} es un término aleatorio o un término de error, y γ_{jk} denota los *efectos de la interacción* renglón-columna (o tratamiento-bloque) que se conocen simplemente como *interacciones*. Se tienen las restricciones

$$\sum_j \alpha_j = 0 \qquad \sum_k \beta_k = 0 \qquad \sum_j \gamma_{jk} = 0 \qquad \sum_k \gamma_{jk} = 0 \tag{45}$$

y se supone que las X_{jkl} están distribuidas de manera normal con media μ y varianza σ^2.

Como antes, la variación V de todos los datos puede dividirse en variaciones debidas a los renglones V_R, variaciones debidas a las columnas V_C, interacciones V_I y un error aleatorio o residual V_E:

$$V = V_R + V_C + V_I + V_E \tag{46}$$

donde

$$V = \sum_{j,k,l} (X_{jkl} - \bar{X})^2 \tag{47}$$

$$V_R = bc\sum_{j=1}^{a} (\bar{X}_{j..} - \bar{X})^2 \tag{48}$$

$$V_C = ac\sum_{k=1}^{b} (\bar{X}_{.k.} - \bar{X})^2 \tag{49}$$

$$V_I = c\sum_{j,k} (\bar{X}_{jk.} - \bar{X}_{j..} - \bar{X}_{.k.} + \bar{X})^2 \tag{50}$$

$$V_E = \sum_{j,k,l} (X_{jkl} - \bar{X}_{jk.})^2 \tag{51}$$

En estos resultados, los puntos que aparecen en los subíndices tienen significados análogos a los dados antes; así, por ejemplo,

$$\bar{X}_{j..} = \frac{1}{bc}\sum_{k,l} X_{jkl} = \frac{1}{b}\sum_k \bar{X}_{jk.} \tag{52}$$

El valor esperado de las variaciones se encuentra como antes. Empleando, para cada fuente de variación, el número que le corresponde de grados de libertad, se puede elaborar una tabla para el análisis de varianza como la que se muestra

en la tabla 16.6. Los cocientes F que aparecen en la última columna de la tabla 16.6 se usan para probar las hipótesis nulas:

$H_0^{(1)}$: todas las medias de los tratamientos (renglones) son iguales; es decir, $\alpha_j = 0$.

$H_0^{(2)}$: todas las medias de los bloques (columnas) son iguales; es decir, $\beta_k = 0$.

$H_0^{(3)}$: entre tratamientos y bloques no hay interacción; es decir, $\gamma_{jk} = 0$.

Tabla 16.6

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, V_R	$a-1$	$\hat{S}_R^2 = \dfrac{V_R}{a-1}$	$\hat{S}_R^2/\hat{S}_E^2$ con $a-1$ y $ab(c-1)$ grados de libertad
Entre bloques, V_C	$b-1$	$\hat{S}_C^2 = \dfrac{V_C}{b-1}$	$\hat{S}_C^2/\hat{S}_E^2$ con $b-1$ y $ab(c-1)$ grados de libertad
Interacción, V_I	$(a-1)(b-1)$	$\hat{S}_I^2 = \dfrac{V_I}{(a-1)(b-1)}$	$\hat{S}_I^2/\hat{S}_E^2$ con $(a-1)(b-1)$ y $ab(c-1)$ grados de libertad
Residual o aleatoria, V_E	$ab(c-1)$	$\hat{S}_E^2 = \dfrac{V_E}{ab(c-1)}$	
Total, V	$abc-1$		

Desde un punto de vista práctico, hay que decidir primero si $H_0^{(3)}$ puede o no ser rechazada a nivel de significancia apropiado usando el F-cociente $\hat{S}_I^2/\hat{S}_E^2$ de la tabla 16.6. Pueden presentarse dos casos:

1. **$H_0^{(3)}$ no puede ser rechazada.** En este caso se concluye que las interacciones no son muy grandes. Entonces, se pueden probar $H_0^{(1)}$ y $H_0^{(2)}$ empleando, respectivamente, los F-cocientes $\hat{S}_R^2/\hat{S}_E^2$ y $\hat{S}_C^2/\hat{S}_E^2$ como se muestra en la tabla 16.6. Algunos especialistas en estadística recomiendan que en este caso se junten las variaciones y se use el total $V_I + V_E$ dividiéndolo entre la correspondiente suma de grados de libertad $(a-1)(b-1) + ab(c-1)$ y usando, en la prueba F, este valor en lugar de $\hat{S}_E^2$.

2. **$H_0^{(3)}$ puede ser rechazada.** En este caso, se concluye que las interacciones son significativamente grandes. Entonces, las diferencias entre los factores sólo serán importantes si son grandes en comparación con estas interacciones. A esto se debe que muchos especialistas en estadística recomienden probar $H_0^{(1)}$ y $H_0^{(2)}$ empleando los F-cocientes $\hat{S}_R^2/\hat{S}_I^2$ y $\hat{S}_C^2/\hat{S}_I^2$ en lugar de los dados en la tabla 16.6. Aquí también se usará este procedimiento alternativo.

El análisis de varianza con replicación puede realizarse más fácilmente sumando primero los valores de las replicaciones correspondientes a un tratamiento (renglón) y a un bloque (columna). Con esto se obtiene una tabla de dos factores con entrada sencilla, que se puede analizar como la tabla 16.5. Este procedimiento se ilustra en el problema 16.16.

DISEÑO EXPERIMENTAL

Las técnicas de análisis de varianza vistas antes se emplean una vez que se han obtenido los resultados de un experimento. Sin embargo, con objeto de obtener tanta información como sea posible, es necesario que primero se planee

cuidadosamente el experimento; a esto se le conoce como *diseño del experimento*. Los siguientes son algunos ejemplos importantes de diseños de experimentos:

1. **Aleatorización completa.** Supóngase que se tiene un experimento agrícola como el del ejemplo 1. Para diseñar este experimento se puede dividir la tierra en $4 \times 4 = 16$ parcelas (como se indica en la figura 16-1 mediante los cuadrados, aunque puede emplearse cualquier otra figura) y asignar cada tratamiento (indicados por las letras *A*, *B*, *C* y *D*) a cuatro bloques elegidos en forma completamente aleatoria. El propósito de la aleatorización es eliminar diversas fuentes de error, por ejemplo, la fertilidad del suelo.

D	*A*	*C*	*C*
B	*D*	*B*	*A*
D	*C*	*B*	*D*
A	*B*	*C*	*A*

Figura 16-1 Aleatorización completa.

Bloques	Tratamientos			
I	*C*	*B*	*A*	*D*
II	*A*	*B*	*D*	*C*
III	*B*	*C*	*D*	*A*
IV	*A*	*D*	*C*	*B*

Figura 16-2 Bloques aleatorizados.

	Factor 1			
Factor 2	*D*	*B*	*C*	*A*
	B	*D*	*A*	*C*
	C	*A*	*D*	*B*
	A	*C*	*B*	*D*

Figura 16-3 Cuadrado latino.

B_γ	A_β	D_δ	C_α
A_δ	B_α	C_γ	D_β
D_α	C_δ	B_β	A_γ
C_β	D_γ	A_α	B_δ

Figura 16-4 Cuadrado grecolatino.

2. **Bloques aleatorizados.** Cuando se necesita todo un conjunto de tratamientos para cada bloque, como en el ejemplo 2, los tratamientos *A*, *B*, *C* y *D* se introducen en orden aleatorio en cada uno de los bloques I, II, III y IV (es decir, en los renglones de la figura 16-2), y por esta razón a los bloques se les llama *bloques aleatorizados*. Este tipo de diseño se emplea para controlar *una fuente de error o variabilidad*: a saber, la diferencia entre los bloques.
3. **Cuadrados latinos.** Para algunos fines es necesario controlar al mismo tiempo *dos fuentes de error o de variabilidad*, como las diferencias entre los renglones y las diferencias entre las columnas. Por ejemplo, en el experimento del ejemplo 1, los errores en los diferentes renglones y columnas pueden deberse a variaciones en la fertilidad del suelo en distintos lugares del terreno. En tales casos es necesario que cada tratamiento aparezca una vez en cada renglón y una vez en cada columna, como en la figura 16-3. A esta distribución se le llama *cuadrado latino* debido a que se emplean las letras *A*, *B*, *C* y *D*.
4. **Cuadrados grecolatinos.** Cuando es necesario controlar *tres fuentes de error o de variabilidad* se emplea un *cuadrado grecolatino*, como el que se muestra en la figura 16-4. Estos cuadrados son, en esencia, dos cuadrados latinos superpuestos uno sobre otro, usando las letras latinas *A*, *B*, *C* y *D* para uno de los cuadrados y las letras griegas α, β, γ y δ para el otro. Un requerimiento adicional por satisfacer es que cada letra griega debe usarse una y sólo una vez con cada letra latina; si se satisface esta condición se dice que el cuadrado es *ortogonal*.

PROBLEMAS RESUELTOS

CLASIFICACIÓN EN UN SENTIDO O EXPERIMENTOS CON UN FACTOR

16.1 Probar que $V = V_W + V_B$; es decir,

$$\sum_{j,k}(X_{jk} - \bar{X})^2 = \sum_{j,k}(X_{jk} - \bar{X}_{j.})^2 + \sum_{j,k}(\bar{X}_{j.} - \bar{X})^2$$

SOLUCIÓN

Se tiene
$$X_{jk} - \bar{X} = (X_{jk} - \bar{X}_{j.}) + (\bar{X}_{j.} - \bar{X})$$

Elevando al cuadrado y sumando sobre j y k, se obtiene

$$\sum_{j,k}(X_{jk} - \bar{X})^2 = \sum_{j,k}(X_{jk} - \bar{X}_{j.})^2 + \sum_{j,k}(\bar{X}_{j.} - \bar{X})^2 + 2\sum_{j,k}(X_{jk} - \bar{X}_{j.})(\bar{X}_{j.} - \bar{X})$$

Para probar el resultado deseado hay que mostrar que la última suma es cero. Para esto, se procede como sigue:

$$\begin{aligned}\sum_{j,k}(X_{jk} - \bar{X}_{j.})(\bar{X}_{j.} - \bar{X}) &= \sum_{j=1}^{a}(\bar{X}_{j.} - \bar{X})\left[\sum_{k=1}^{b}(X_{jk} - \bar{X}_{j.})\right] \\ &= \sum_{j=1}^{a}(\bar{X}_{j.} - \bar{X})\left[\left(\sum_{k=1}^{b}X_{jk}\right) - b\bar{X}_{j.}\right] = 0\end{aligned}$$

ya que
$$\bar{X}_{j.} = \frac{1}{b}\sum_{k=1}^{b}X_{jk}$$

16.2 Empleando la notación de la página 362, verificar que: *a*) $T = ab\bar{X}$, *b*) $T_{j.} = b\bar{X}_{j.}$ y *c*) $\sum_j T_{j.} = ab\bar{X}$.

SOLUCIÓN

a)
$$T = \sum_{j,k}X_{jk} = ab\left(\frac{1}{ab}\sum_{j,k}X_{jk}\right) = ab\bar{X}$$

b)
$$T_{j.} = \sum_{k}X_{jk} = b\left(\frac{1}{b}\sum_{k}X_{jk}\right) = b\bar{X}_{j.}$$

c) Como $T_{j.} = \sum_k X_{jk}$, de acuerdo con el inciso *a*) se tiene

$$\sum_{j}T_{j.} = \sum_{j}\sum_{k}X_{jk} = T = ab\bar{X}$$

16.3 Verificar las fórmulas abreviadas (*10*), (*11*) y (*12*) de este capítulo.

SOLUCIÓN

Se tiene
$$\begin{aligned}V = \sum_{j,k}(X_{jk} - \bar{X})^2 &= \sum_{j,k}(X_{jk}^2 - 2\bar{X}X_{jk} + \bar{X}^2) \\ &= \sum_{j,k}X_{jk}^2 - 2\bar{X}\sum_{j,k}X_{jk} + ab\bar{X}^2 \\ &= \sum_{j,k}X_{jk}^2 - 2\bar{X}(ab\bar{X}) + ab\bar{X}^2 \\ &= \sum_{j,k}X_{jk}^2 - ab\bar{X}^2 \\ &= \sum_{j,k}X_{jk}^2 - \frac{T^2}{ab}\end{aligned}$$

empleando el problema 16.2*a*) para el tercero y último renglones anteriores. De igual manera,

$$\begin{aligned} V_B &= \sum_{j,k} (\bar{X}_{j.} - \bar{X})^2 = \sum_{j,k} (\bar{X}_{j.}^2 - 2\bar{X}\bar{X}_{j.} + \bar{X}^2) \\ &= \sum_{j,k} \bar{X}_{j.}^2 - 2\bar{X} \sum_{j,k} \bar{X}_{j.} + ab\bar{X}^2 \\ &= \sum_{j,k} \left(\frac{T_{j.}}{b}\right)^2 - 2\bar{X} \sum_{j,k} \frac{T_{j.}}{b} + ab\bar{X}^2 \\ &= \frac{1}{b^2} \sum_{j=1}^{a} \sum_{k=1}^{b} T_{j.}^2 - 2\bar{X}(ab\bar{X}) + ab\bar{X}^2 \\ &= \frac{1}{b} \sum_{j=1}^{a} T_{j.}^2 - ab\bar{X}^2 \\ &= \frac{1}{b} \sum_{j=1}^{a} T_{j.}^2 - \frac{T^2}{ab} \end{aligned}$$

empleando el problema 16.2*b*) para el tercer renglón y el problema 16.2*a*) para el último renglón. Por último, la ecuación (*12*) se obtiene a partir de que $V = V_W + V_B$ o bien $V_W = V - V_B$.

16.4 La tabla 16.7 muestra los rendimientos, en bushels por acre, de cierta variedad de trigo cultivado en un tipo especial de suelo tratado con los agentes químicos *A*, *B* o *C*. Encontrar: *a*) el rendimiento medio con los distintos tratamientos, *b*) la gran media de todos los tratamientos, *c*) la variación total, *d*) la variación entre los tratamientos y *e*) la variación dentro de los tratamientos. Utilizar el método largo. *f*) Proporcionar el análisis de EXCEL para los datos que se muestran en la tabla 16.7.

Tabla 16.7

A	48	49	50	49
B	47	49	48	48
C	49	51	50	50

Tabla 16.8

3	4	5	4
2	4	3	3
4	6	5	5

SOLUCIÓN

Para simplificar los cálculos se puede sustraer una cantidad adecuada, por ejemplo 45, de cada uno de los datos sin que esto afecte los valores de las variaciones. Así se obtienen los datos de la tabla 16.8.

a) Las medias de tratamiento (renglón) en la tabla 16.8 son, respectivamente,

$$\bar{X}_{1.} = \tfrac{1}{4}(3+4+5+4) = 4 \qquad \bar{X}_{2.} = \tfrac{1}{4}(2+4+3+3) = 3 \qquad \bar{X}_{3.} = \tfrac{1}{4}(4+6+5+5) = 5$$

Y los rendimientos medios, que se obtienen sumando 45 a estos valores, son 49, 48 y 50 bushels por acre, respectivamente, para *A*, *B* y *C*.

b) La gran media de todos los tratamientos es

$$\bar{X} = \tfrac{1}{12}(3+4+5+4+2+4+3+3+4+6+5+5) = 4$$

Por lo tanto, la gran media del conjunto de los datos originales es $45 + 4 = 49$ bushels por acre.

c) La variación total es

$$\begin{aligned} V = \sum_{j,k} (X_{jk} - X)^2 &= (3-4)^2 + (4-4)^2 + (5-4)^2 + (4-4)^2 + (2-4)^2 + (4-4)^2 \\ &\quad + (3-4)^2 + (3-4)^2 + (4-4)^2 + (6-4)^2 + (5-4)^2 + (5-4)^2 = 14 \end{aligned}$$

d) La variación entre tratamientos es

$$V_B = b\sum_j (\bar{X}_{j.} - \bar{X})^2 = 4[(4-4)^2 + (3-4)^2 + (5-4)^2] = 8$$

e) La variación dentro de los tratamientos es

$$V_W = V - V_B = 14 - 8 = 6$$

Otro método

$$V_W = \sum_{j,k} (X_{jk} - X_{j.})^2 = (3-4)^2 + (4-4)^2 + (5-4)^2 + (4-4)^2 + (2-3)^2 + (4-3)^2$$
$$+ (3-3)^2 + (3-3)^2 + (4-5)^2 + (6-5)^2 + (5-5)^2 + (5-5)^2 = 6$$

Nota: La tabla 16.9 es para el análisis de varianza de los problemas 16.4, 16.5 y 16.6.

Tabla 16.9

Variación	Grados de libertad	Cuadrado medio	*F*
Entre tratamientos $V_B = 8$	$a - 1 = 2$	$\hat{S}_B^2 = \frac{8}{2} = 4$	$\frac{\hat{S}_B^2}{\hat{S}_W^2} = \frac{4}{2/3} = 6$
Dentro de los tratamientos $V_W = V - V_B = 14 - 8 = 6$	$a(b-1) = (3)(3) = 9$	$\hat{S}_W^2 = \frac{6}{9} = \frac{2}{3}$	con 2 y 9 grados de libertad
Total $V = 14$	$ab - 1 = (3)(4) - 1 = 11$		

f) Empleando EXCEL, la secuencia **Tools** → **Data analysis** → **Anova single factor** da el análisis que se presenta a continuación. El valor *p* indica que $\alpha = 0.05$, las medias de las tres variedades son diferentes.

A	**B**	**C**
48	47	49
49	49	51
50	48	50
49	48	50

Análisis de varianza de un factor

RESUMEN

Grupos	*Cuenta*	*Suma*	*Promedio*	*Varianza*
A	4	196	49	0.666667
B	4	192	48	0.666667
C	4	200	50	0.666667

ANÁLISIS DE VARIANZA

Origen de las variaciones	*SS*	*df*	*MS*	*F*	*Valor p*
Entre grupos	8	2	4	6	0.022085
Dentro de los grupos	6	9	0.666667		
Total	14	11			

La figura 16-5 muestra una gráfica de puntos de MINITAB dando los rendimientos de las tres variedades de trigo. La figura 16-6 muestra una gráfica de caja de MINITAB dando los rendimientos de las tres variedades de trigo. El análisis de EXCEL y las gráficas de MINITAB indican que la variedad *C* supera significativamente los rendimientos de la variedad *B*.

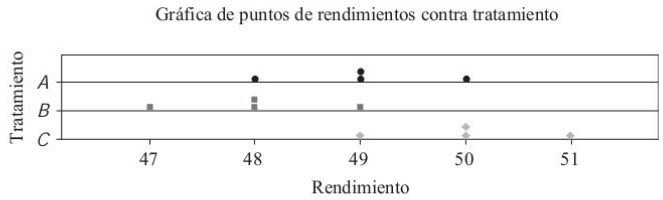


Figura 16-5 MINITAB, gráfica de puntos de los rendimientos de las tres variedades de trigo.

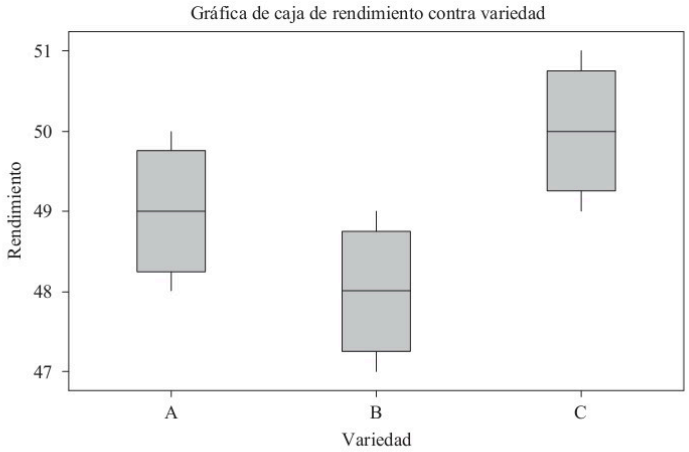


Figura 16-6 MINITAB, gráfica de caja de los rendimientos de las tres variedades de trigo.

16.5 Volver al problema 16.4, encontrar una estimación insesgada de la varianza poblacional σ^2 a partir de: *a*) la variación entre tratamientos bajo la hipótesis nula de medias de tratamiento iguales y *b*) la variación dentro de los tratamientos. *c*) Consultar los resultados de EXCEL dados en la solución del problema 16.4, localizar las estimaciones de las varianzas calculadas en los incisos *a*) y *b*).

SOLUCIÓN

a)
$$\hat{S}_B^2 = \frac{V_B}{a-1} = \frac{8}{3-1} = 4$$

b)
$$\hat{S}_W^2 = \frac{V_W}{a(b-1)} = \frac{6}{3(4-1)} = \frac{2}{3}$$

c) La estimación de varianza $\hat{S}_B^2$, en los resultados de EXCEL, es MS entre grupos y es 4, que es igual al valor encontrado. La estimación de $\hat{S}_W^2$, en los resultados de EXCEL, es MS dentro de los grupos y es 0.666667, que es igual al valor encontrado.

16.6 Dados los datos del problema 16.4, a los niveles de significancia: *a*) 0.05 y *b*) 0.01, ¿puede rechazarse la hipótesis nula de medias iguales? *c*) Consultar los resultados de EXCEL dados en la solución del problema 16.4, para probar la hipótesis nula de varianzas iguales.

SOLUCIÓN

Se tiene $$F = \frac{\hat{S}_B^2}{\hat{S}_W^2} = \frac{4}{2/3} = 6$$

con $a - 1 = 3 - 1$ grados de libertad y $a(b-1) = 3(4-1) = 9$ grados de libertad.

a) En el apéndice V, para $\nu_1 = 2$ y $\nu_2 = 9$, se encuentra que $F_{.95} = 4.26$. Como $F = 6 > F_{.95}$, la hipótesis nula de medias iguales puede rechazarse al nivel 0.05.

b) En el apéndice VI, para $\nu_1 = 2$ y $\nu_2 = 9$, se encuentra que $F_{.99} = 8.02$. Como $F = 6 < F_{.99}$, la hipótesis nula de medias iguales no se puede rechazar al nivel 0.01.

c) Consultando los resultados de EXCEL dados en el problema 16.4, se encuentra que el valor F es 6 y el valor p es 0.022. Por lo tanto, el menor nivel de significancia predeterminado al que puede rechazarse la hipótesis nula es 0.022. De manera que la hipótesis nula se rechazará al nivel de significancia 0.05, pero no al nivel de significancia 0.01.

16.7 Dados los datos del problema 16.4, emplear las fórmulas abreviadas (*10*), (*11*) y (*12*) para obtener: *a*) la variación total, *b*) la variación entre los tratamientos y *c*) la variación dentro de los tratamientos. Además, utilizar MINITAB con los datos, a los que se les restó 45 a cada valor, para obtener la tabla del análisis de varianza.

SOLUCIÓN

Conviene ordenar los datos como en la tabla 16.10.

Tabla 16.10

		$T_{j\cdot}$	$T_{j\cdot}^2$
A	3 4 5 4	16	256
B	2 4 3 3	12	144
C	4 6 5 5	20	400
	$\sum_{j,k} X_{jk}^2 = 206$	$T = \sum_j T_{j\cdot} = 48$	$\sum_j T_{j\cdot}^2 = 800$

a) Usando la fórmula (*10*), se tiene

$$\sum_{j,k} X_{jk}^2 = 9 + 16 + 25 + 16 + 4 + 16 + 9 + 9 + 16 + 36 + 25 + 25 = 206$$

y $$T = 3 + 4 + 5 + 4 + 2 + 4 + 3 + 3 + 4 + 6 + 5 + 5 = 48$$

Por lo tanto, $$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} = 206 - \frac{(48)^2}{(3)(4)} = 206 - 192 = 14$$

b) Los totales (suma) de los renglones son

$$T_{1\cdot} = 3 + 4 + 5 + 4 = 16 \qquad T_{2\cdot} = 2 + 4 + 3 + 3 = 12 \qquad T_{3\cdot} = 4 + 6 + 5 + 5 = 20$$

y $$T = 16 + 12 + 20 = 48$$

Por lo tanto, empleando la fórmula (*11*), se tiene

$$V_B = \frac{1}{b}\sum_j T_{j\cdot}^2 - \frac{T^2}{ab} = \frac{1}{4}(16^2 + 12^2 + 20^2) - \frac{(48)^2}{(3)(4)} = 200 - 192 = 8$$

c) Empleando la fórmula (*12*), se tiene

$$V_W = V - V_B = 14 - 8 = 6$$

Estos resultados coinciden con los obtenidos en el problema 16.4 y se procede como antes.

Con la secuencia **Stat** → **Anova** → **Oneway** se obtiene el resultado siguiente. Obsérvense las diferencias en la terminología empleada. A la variación dentro de los tratamientos se le llama en EXCEL `Dentro de los grupos` y en MINITAB `Error`. A la variación entre tratamientos en EXCEL se le llama `Entre los grupos` y en MINITAB `Factor`. El usuario debe acostumbrarse a las diferentes terminologías empleadas en los diversos paquetes de software.

One-way ANOVA: A, B, C

```
Source   DF      SS     MS     F      P
Factor    2   8.000  4.000  6.00  0.022
Error     9   6.00   0.667
Total    11  14.000

S=0.8165      R-Sq=57.14%      R-Sq(adj)=47.62%
```

16.8 Una empresa quiere comprar una de cinco máquinas *A*, *B*, *C*, *D* o *E*. En un experimento destinado a probar si hay diferencia en el rendimiento de estas máquinas, uno de cada cinco operadores experimentados trabaja durante la misma cantidad de tiempo en cada máquina. En la tabla 16.11 se muestra la cantidad de unidades producidas con cada máquina. A los niveles de significancia: *a*) 0.05 y *b*) 0.01, probar la hipótesis de que no hay diferencia entre las máquinas. *c*) Proporcionar la solución de STATISTIX a este problema, y empleando el método del valor *p*, probar la hipótesis de que no hay diferencia entre las máquinas. Usar $\alpha = 0.05$.

Tabla 16.11

A	68	72	77	42	53
B	72	53	63	53	48
C	60	82	64	75	72
D	48	61	57	64	50
E	64	65	70	68	53

Tabla 16.12

						$T_{j\cdot}$	$T_{j\cdot}^2$
A	8	12	17	−18	−7	12	144
B	12	−7	3	−7	−12	−11	121
C	0	22	4	15	12	53	2 809
D	−12	1	−3	4	−10	−20	400
E	4	5	10	8	−7	20	400
	$\sum X_{jk}^2 = 2\,658$					54	3 874

SOLUCIÓN

A cada dato se le resta un número adecuado, por ejemplo 60, y se obtiene la tabla 16.12. Entonces

$$V = 2\,658 - \frac{(54)^2}{(5)(5)} = 2\,658 - 116.64 = 2\,541.36$$

y

$$V_B = \frac{3\,874}{5} - \frac{(54)^2}{(5)(4)} = 774.8 - 116.64 = 658.16$$

Ahora se elabora la tabla 16.13. Para 4 y 20 grados de libertad, se tiene $F_{.95} = 2.87$. De esta manera, al nivel 0.05 no se puede rechazar la hipótesis nula y, por lo tanto, tampoco al nivel 0.01.

Con la secuencia **Statistics** → **One, two, multi-sample tests** → **One-way Anova** se obtiene el resultado siguiente.

```
Statistix 8.
One-Way AOV for: A B C D E

Source      DF          SS          MS          F          P
Between      4      658.16     164.540       1.75     0.1792
Within      20     1883.20      94.160
Total       24     2541.36

Grand Mean 62.160      CV 15.61

Variable      Mean
A           62.400
B           57.800
C           70.600
D           56.000
E           64.000
```

El valor p es 0.1792. No hay diferencia significativa entre las medias poblacionales.

Tabla 16.13

Variación	Grados de libertad	Cuadrado medio	F
Entre tratamientos, $V_B = 658.2$	$a - 1 = 4$	$\hat{S}_B^2 = \dfrac{658.2}{4} = 164.5$	$F = \dfrac{164.55}{94.16} = 1.75$
Dentro de tratamientos, $V_W = 1\,883.2$	$a(b-1) = (5)(4) = 20$	$\hat{S}_W^2 = \dfrac{1\,883.2}{20} = 94.16$	
Total, $V = 2\,514.4$	$ab - 1 = 24$		

MODIFICACIONES PARA NÚMEROS DISTINTOS DE OBSERVACIONES

16.9 En la tabla 16.14 se presentan las duraciones, en horas, de muestras de tres diferentes tipos de cinescopios producidos por una empresa. Usando el método largo, a los niveles de significancia: *a*) 0.05 y *b*) 0.01, determinar si hay alguna diferencia entre los tres tipos de cinescopios.

Tabla 16.14

Muestra 1	407	411	409		
Muestra 2	404	406	408	405	402
Muestra 3	410	408	406	408	

SOLUCIÓN

Para facilitar los cálculos se resta de cada dato un número apropiado, por ejemplo 400, obteniendo así la tabla 16.15. En esta tabla se dan los totales de los renglones, las medias muestrales (o grupales) y la gran media. De esta manera se tiene

$$V = \sum_{j,k} (X_{jk} - \bar{X})^2 = (7-7)^2 + (11-7)^2 + \cdots + (8-7)^2 = 72$$

$$V_B = \sum_{j,k} (\bar{X}j. - \bar{X})^2 = \sum_{j} N_j(\bar{X}_{j.} - \bar{X})^2 = 3(9-7)^2 + 5(7-5)^2 + 4(8-7)^2 = 36$$

$$V_W = V - V_B = 72 - 36 = 36$$

V_W también puede obtenerse directamente observando que es igual a

$$(7-9)^2 + (11-9)^2 + (9-9)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2 + (5-5)^2$$
$$+ (2-5)^2 + (10-8)^2 + (8-8)^2 + (6-8)^2 + (8-8)^2$$

Tabla 16.15

						Total	Media
Muestra 1	7	11	9			27	9
Muestra 2	4	6	8	5	2	25	5
Muestra 3	10	8	6	8		32	8
	$\bar{X}$ = gran media = $\frac{84}{12} = 7$						

Los datos pueden resumirse como en la tabla 16.16, la tabla para el análisis de varianza. Para 2 y 9 grados de libertad, en el apéndice V se encuentra que $F_{.95} = 4.26$ y en el apéndice VI que $F_{.99} = 8.02$. Por lo tanto, la hipótesis de que las medias son iguales (es decir, que no hay diferencia entre los tres tipos de cinescopios) puede rechazarse al nivel de significancia 0.05, pero no al nivel de significancia 0.01.

Tabla 16.16

Variación	Grados de libertad	Cuadrado medio	F
$V_B = 36$	$a - 1 = 2$	$\hat{S}_B^2 = \frac{36}{2} = 18$	$\frac{\hat{S}_B^2}{\hat{S}_W^2} = \frac{18}{4}$
$V_W = 36$	$N - a = 9$	$\hat{S}_W^2 = \frac{36}{9} = 4$	$= 4.5$

16.10 Resolver el problema 16.9 empleando las fórmulas abreviadas (*24*), (*25*) y (*26*). Además, proporcionar la solución al problema empleando SAS.

SOLUCIÓN

De acuerdo con la tabla 16.15, se tiene $N_1 = 3$, $N_2 = 5$, $N_3 = 4$, $N = 12$, $T_{1.} = 27$, $T_{2.} = 25$, $T_{3.} = 32$ y $T = 84$. Por lo tanto, se tiene

$$V = \sum_{j,k} X_{Jk}^2 - \frac{T^2}{N} = 7^2 + 11^2 + \cdots + 6^2 + 8^2 - \frac{(84)^2}{12} = 72$$

$$V_B = \sum_{j} \frac{T_{J.}^2}{N_j} - \frac{T^2}{N} = \frac{(27)^2}{3} + \frac{(25)^2}{5} + \frac{(32)^2}{4} - \frac{(84)^2}{12} = 36$$

$$V_W = V - V_B = 36$$

Empleando estos valores, el análisis de varianza procede entonces como en el problema 16.9.

Empleando SAS, con la secuencia **Statistics** → **ANOVA** → **Oneway ANOVA** se obtienen los resultados siguientes.

```
The ANOVA Procedure
                    Class Level Information
                  Class      Levels    Values
                  Sample_         3     1 2 3

                Number of Observations Read   12
                Number of Observations Used   12

                    The ANOVA Procedure
Dependent Variable: lifetime

Source                DF       Sum of Squares   Mean Square     F value          Pr > F
Model                  2          36.00000000   18.00000000        4.50          0.0442
Error                  9          36.00000000    4.00000000
Corrected Total       11          72.00000000

                                    R-Square     coeff Var    Root MSE   lifetime Mean
                                    0.500000      0.491400    2.000000        407.0000

Source                DF             Anova SS   Mean Square     F Value          Pr > F
Sample_                2          36.00000000   18.00000000        4.50          0.0442
```

Obsérvese que en SAS a la variación entre tratamientos se le llama `model` (modelo) y a la variación dentro de los tratamientos se le dice `error`. Al estadístico de prueba se le llama valor F y es igual a 4.50. El valor p es $Pr > F$ y es igual a 0.0442. A $\alpha = 0.05$ se declarará que las duraciones no son iguales.

CLASIFICACIÓN EN DOS SENTIDOS O EXPERIMENTOS CON DOS FACTORES

16.11 En la tabla 16.17 se presenta la producción por acre en cuatro cultivos diferentes empleando tres tipos diferentes de fertilizantes. Usando el método largo, determinar, al nivel de significancia 0.01, si hay diferencias en la producción por acre: *a*) debidas a los fertilizantes y *b*) debidas a los cultivos. *c*) Proporcionar la solución que da MINITAB a este experimento de dos factores.

SOLUCIÓN

Como se muestra en la tabla 16.18, se calculan los totales de los renglones, las medias de los renglones, los totales de las columnas, las medias de las columnas, el gran total y la gran media. Según esta tabla se obtiene:

La variación de las medias de los renglones respecto a la gran media es

$$V_R = 4[(6.2 - 6.8)^2 + (8.3 - 6.8)^2 + (5.9 - 6.8)^2] = 13.68$$

La variación de las medias de las columnas respecto a la gran media es

$$V_C = 3[(6.4 - 6.8)^2 + (7.0 - 6.8)^2 + (7.5 - 6.8)^2 + (6.3 - 6.8)^2] = 2.82$$

Tabla 16.17

	Cultivo I	Cultivo II	Cultivo III	Cultivo IV
Fertilizante *A*	4.5	6.4	7.2	6.7
Fertilizante *B*	8.8	7.8	9.6	7.0
Fertilizante *C*	5.9	6.8	5.7	5.2

Tabla 16.18

	Cultivo I	Cultivo II	Cultivo III	Cultivo IV	Total del renglón	Media del renglón
Fertilizante *A*	4.5	6.4	7.2	6.7	24.8	6.2
Fertilizante *B*	8.8	7.8	9.6	7.0	33.2	8.3
Fertilizante *C*	5.9	6.8	5.7	5.2	23.6	5.9
Total de la columna	19.2	21.0	22.5	18.9	Gran total = 81.6	
Media de la columna	6.4	7.0	7.5	6.3	Gran media = 6.8	

La variación total es

$$V = (4.5 - 6.8)^2 + (6.4 - 6.8)^2 + (7.2 - 6.8)^2 + (6.7 - 6.8)^2$$
$$+ (8.8 - 6.8)^2 + (7.8 - 6.8)^2 + (9.6 - 6.8)^2 + (7.0 - 6.8)^2$$
$$+ (5.9 - 6.8)^2 + (6.8 - 6.8)^2 + (5.7 - 6.8)^2 + (5.2 - 6.8)^2 = 23.08$$

La variación aleatoria es

$$V_E = V - V_R - V_C = 6.58$$

Esto conduce al análisis de varianza de la tabla 16.19.

Tabla 16.19

Variación	Grados de libertad	Cuadrado medio	F
$V_R = 13.68$	2	$\hat{S}_R^2 = 6.84$	$\hat{S}_R^2/\hat{S}_E^2 = 6.24$ con 2 y 6 grados de libertad
$V_C = 2.82$	3	$\hat{S}_C^2 = 0.94$	$\hat{S}_C^2/\hat{S}_E^2 = 0.86$ con 3 y 6 grados de libertad
$V_E = 6.58$	6	$\hat{S}_E^2 = 1.097$	
$V = 23.08$	11		

a) Al nivel de significancia 0.05 con 2 y 6 grados de libertad, $F_{95} = 5.14$. Entonces, como 6.24 > 5.14, se puede rechazar la hipótesis de que las medias de los renglones sean iguales y concluir que al nivel de significancia 0.05 existe, en la producción, una diferencia significativa debida a los fertilizantes.

b) Como el valor F correspondiente a las diferencias en las medias de las columnas es menor que 1, se concluye que debido a los cultivos no hay diferencia significativa en la producción.

c) Primero se da la estructura que deben tener los datos en la hoja de cálculo de MINITAB, y a continuación el análisis de MINITAB para este experimento de dos factores.

```
 Row   Crop    Fertilizer    Yield
   1      1             1      4.5
   2      1             2      8.8
   3      1             3      5.9
   4      2             1      6.4
   5      2             2      7.8
   6      2             3      6.8
   7      3             1      7.2
   8      3             2      9.6
   9      3             3      5.7
  10      4             1      6.7
  11      4             2      7.0
  12      4             3      5.2

MTB > Twoway 'Yield' 'Crop' 'Fertilizer';
SUBC > Means 'Crop' 'Fertilizer'.
```

Two-way Analysis of Variance

```
Analysis of Variance for Yield
Source       DF        SS        MS         F         P
Crop          3      2.82      0.94      0.86     0.512
Fertiliz      2     13.68      6.84      6.24     0.034
Error         6      6.58      1.10
Total        11     23.08

                    Individual 95% CI
Crop       Mean     --+---------+---------+---------+---------
1          6.40      (---------------*---------------)
2          7.00                (---------------*---------------)
3          7.50                    (---------------*---------------)
4          6.30     (---------------*---------------)
                    --+---------+---------+---------+---------
                    5.00      6.00      7.00      8.00

                    Individual 95% CI
Fertiliz   Mean     --+---------+---------+---------+---------
1          6.20         (----------*----------)
2          8.30                              (----------*----------)
3          5.90     (----------*----------)
                    --+---------+---------+---------+---------
                      4.80      6.00      7.20      8.40
```

La estructura de los datos en la hoja de cálculo debe corresponder exactamente a la estructura de los datos en la tabla 16.17. El primer renglón, `1 1 4.5`, corresponde a Cultivo 1, Fertilizante 1 y Rendimiento 4.5; el segundo renglón, `1 2 8.8`, corresponde a Cultivo 1, Fertilizante 2 y Rendimiento 8.8, etc. Un error frecuente al usar software para estadística es que en la hoja de cálculo se dé una estructura incorrecta de los datos. Hay que asegurarse de que los datos dados en una tabla como la 16.17 y la estructura de los datos en la hoja de cálculo se correspondan uno a uno. Obsérvese que la tabla para el análisis de varianza en dos sentidos, dada en los resultados de MINITAB, contiene la información de la tabla 16.19. Los valores p que aparecen en los resultados de MINITAB permiten al investigador probar la hipótesis de interés sin tener que consultar las tablas de la distribución F para hallar los valores críticos. El valor p para los cultivos es 0.512. Éste es el nivel de significancia mínimo al que se puede rechazar que haya diferencia en la producción media de los cultivos. Las producciones medias de los cuatro cultivos no son estadísticamente significativas a 0.05 o bien 0.01. El valor p para los fertilizantes es 0.034. Esto indica que las producciones medias con los tres fertilizantes son estadísticamente diferentes a 0.05 pero no a 0.01.

Los intervalos de confianza para las medias de los cuatro cultivos dados en los resultados de MINITAB refuerzan la conclusión de que no hay diferencia en las producciones medias de los cuatro diferentes cultivos. Los intervalos de confianza para los tres fertilizantes indican que posiblemente con el fertilizante B se obtenga una producción media más alta que con cualquiera de los fertilizantes A o bien C.

16.12 Usar la fórmula de cálculo abreviada para resolver el problema 16.11. Además, proporcionar la solución a este problema empleando SPSS.

SOLUCIÓN

De acuerdo con la tabla 16.18, se tiene

$$\sum_{j,k} X_{jk}^2 = (4.5)^2 + (6.4)^2 + \cdots + (5.2)^2 = 577.96$$

$$T = 24.8 + 33.2 + 23.6 = 81.6$$

$$\sum T_{j.}^2 = (24.8)^2 + (33.2)^2 + (23.6)^2 = 2\,274.24$$

$$\sum T_{.k}^2 = (19.2)^2 + (21.0)^2 + (22.5)^2 + (18.9)^2 = 1\,673.10$$

Entonces

$$V = \sum_{j,k} X_{jk}^2 - \frac{T^2}{ab} = 577.96 - 554.88 = 23.08$$

$$V_R = \frac{1}{b}\sum T_{j.}^2 - \frac{T^2}{ab} = \frac{1}{4}(2\,274.24) - 554.88 = 13.68$$

$$V_C = \frac{1}{a}\sum T_{.k}^2 - \frac{T^2}{ab} = \frac{1}{3}(1\,673.10) - 554.88 = 2.82$$

$$V_E = V - V_R - V_C = 23.08 - 13.68 - 2.82 = 6.58$$

Lo cual coincide con los resultados del problema 16.11.
Con la secuencia **Analyze** → **General Linear Model** → **Univariate** de SPSS se obtienen los resultados siguientes:

Pruebas de efectos entre temas

Variable dependiente: rendimiento

Origen	Tipo 1: suma de cuadrados	df	Cuadrado medio	F	Sig.
Modelo correcto	16.500[a]	5	3.300	3.009	.106
Intercepto	554.880	1	554.880	505.970	.000
Cultivo	2.820	3	.940	.857	.512
Fertilizante	13.680	2	6.840	6.237	.034
Error	6.580	6	1.097		
Total	577.960	12			
Total corregido	23.080	11			

[a]R cuadrada = .715 (R cuadrada ajustada = .477)

Obsérvese que el estadístico de prueba está dado por F y que para los cultivos el valor F es 0.857 y el correspondiente valor p es 0.512. El valor F para fertilizante es 6.237 y el correspondiente valor p es 0.034. Estos valores corresponden a los valores de la tabla 16.19, así como a los resultados dados por MINITAB en el problema 16.11.

EXPERIMENTOS CON DOS FACTORES CON REPLICACIÓN

16.13 Un fabricante desea determinar la efectividad de cuatro tipos de máquinas (*A*, *B*, *C* y *D*) en la producción de tornillos. Para esto, obtiene la cantidad de tornillos defectuosos producidos por cada máquina durante los días de una semana determinada en cada uno de los dos turnos; los resultados se muestran en la tabla 16.20. Realizar

un análisis de varianza para determinar, al nivel de significancia 0.05, si existe alguna diferencia: *a*) entre las máquinas y *b*) entre los turnos. *c*) Utilizar también MINITAB para realizar el análisis de varianza y probar las diferencias entre las máquinas y entre los turnos usando un valor *p* apropiado.

Tabla 16.20

Máquina	Primer turno					Segundo turno				
	Lunes	Martes	Miércoles	Jueves	Viernes	Lunes	Martes	Miércoles	Jueves	Viernes
A	6	4	5	5	4	5	7	4	6	8
B	10	8	7	7	9	7	9	12	8	8
C	7	5	6	5	9	9	7	5	4	6
D	8	4	6	5	5	5	7	9	7	10

SOLUCIÓN

Los datos también se pueden organizar de manera equivalente, como en la tabla 16.21. En esta tabla se indican los dos factores principales: la máquina y el turno. Obsérvese que se han indicado dos turnos por cada máquina. Los días de la semana pueden considerarse como réplicas (o repeticiones) del desempeño de cada máquina en los dos turnos. La variación total de todos los datos de la tabla 16.21 es

$$V = 6^2 + 4^2 + 5^2 + \cdots + 7^2 + 10^2 - \frac{(268)^2}{40} = 1\,946 - 1\,795.6 = 150.4$$

Tabla 16.21

Factor I: Máquina	Factor II: Turno	Réplicas					Total
		Lunes	Martes	Miércoles	Jueves	Viernes	
A	1	6	4	5	5	4	24
	2	5	7	4	6	8	30
B	1	10	8	7	7	9	41
	2	7	9	12	8	8	44
C	1	7	5	6	5	9	32
	2	9	7	5	4	6	31
D	1	8	4	6	5	5	28
	2	5	7	9	7	10	38
	Total	57	51	54	47	59	268

Con el fin de considerar los dos factores principales (la máquina y el turno), se concentra la atención en la suma de los valores de las réplicas correspondientes a cada combinación de los factores. Éstas se presentan en la tabla 16.22, que es, por lo tanto, una tabla de dos factores con entradas sencillas. La variación total en la tabla 16.22, a la que se le llamará *variación subtotal* V_S, está dada por

$$V_S = \frac{(24)^2}{5} + \frac{(41)^2}{5} + \frac{(32)^2}{5} + \frac{(28)^2}{5} + \frac{(30)^2}{5} + \frac{(44)^2}{5} + \frac{(31)^2}{5} + \frac{(38)^2}{5} - \frac{(268)^2}{40}$$
$$= 1\,861.2 - 1\,795.6 = 65.6$$

La variación entre renglones está dada por

$$V_R = \frac{(54)^2}{10} + \frac{(85)^2}{10} + \frac{(63)^2}{10} + \frac{(66)^2}{10} - \frac{(268)^2}{40} = 1\,846.6 - 1\,795.6 = 51.0$$

Tabla 16.22

Máquina	Primer turno	Segundo turno	Total
A	24	30	54
B	41	44	85
C	32	31	63
D	28	38	66
Total	125	143	268

La variación entre columnas está dada por

$$V_C = \frac{(125)^2}{20} + \frac{(143)^2}{20} - \frac{(268)^2}{40} = 1\,803.7 - 1\,795.6 = 8.1$$

Si de la variación subtotal V_S se resta ahora la suma de las variaciones entre renglones y las variaciones entre columnas ($V_R + V_C$), se obtiene la variación debida a la *interacción* entre renglones y columnas, la cual está dada por

$$V_I = V_S - V_R - V_C = 65.6 - 51.0 - 8.1 = 6.5$$

Por último, se obtiene la variación residual, que puede considerarse como una variación aleatoria o variación por error V_E (siempre que se crea que los diferentes días de la semana no ocasionan diferencia importante), esta variación se encuentra restando la variación subtotal (es decir, la suma de las variaciones de renglón, de columna y de interacción) de la variación total *V*. Esto da

$$V_E = V - (V_R + V_C + V_I) = V - V_S = 150.4 - 65.6 = 84.8$$

Estas variaciones se muestran en la tabla 16.23, la tabla para el análisis de varianza. En esta tabla también se da el número de grados de libertad que corresponde a cada tipo de variación. Por lo tanto, dado que en la tabla 16.22 hay cuatro

Tabla 16.23

Variación	Grados de libertad	Cuadrado medio	*F*
Renglones (máquinas), $V_R = 51.0$	3	$\hat{S}_R^2 = 17.0$	$\frac{17.0}{2.65} = 6.42$
Columnas (turnos), $V_C = 8.1$	1	$\hat{S}_C^2 = 8.1$	$\frac{8.1}{2.65} = 3.06$
Interacción, $V_I = 6.5$	3	$\hat{S}_I^2 = 2.167$	$\frac{2.167}{2.65} = 0.817$
Subtotal, $V_S = 65.6$	7		
Aleatoria o residual, $V_E = 84.8$	32	$\hat{S}_E^2 = 2.65$	
Total, $V = 150.4$	39		

renglones, la variación debida a los renglones tiene $4 - 1 = 3$ grados de libertad, la variación debida a las dos columnas tiene $2 - 1 = 1$ grado de libertad. Para hallar los grados de libertad debidos a la interacción, se observa que en la tabla 16.22 hay ocho entradas; por lo tanto, el total de grados de libertad es $8 - 1 = 7$. Como 3 de estos 7 grados de libertad se deben a los renglones y 1 se debe a las columnas, el resto $[7 - (3 + 1) = 3]$ se debe a la interacción. Dado que en la tabla original 16.21 hay 40 entradas, el total de grados de libertad es $40 - 1 = 39$. De esta manera, los grados de libertad debidos a la variación aleatoria o residual son $39 - 7 = 32$.

Primero debe determinarse si hay alguna interacción significativa. El valor crítico interpolado de la distribución F con 3 y 32 grados de libertad es 2.90. El valor F calculado para la interacción es 0.817 y no es significativo. Entre las máquinas hay una diferencia significativa, ya que el valor F calculado para las máquinas es 6.42 y el valor crítico es 2.90. El valor crítico para los turnos es 4.15. El valor F calculado para los turnos es 3.06. No hay diferencia en los defectos debida a los turnos.

A continuación se muestra la estructura que deben tener los datos en la hoja de cálculo de MINITAB. Compárese la estructura de los datos con los de la tabla 16.21 para ver la relación entre los dos conjuntos de datos.

Row	Machine	Shift	Defects
1	1	1	6
2	1	1	4
3	1	1	5
4	1	1	5
5	1	1	4
6	1	2	5
7	1	2	7
8	1	2	4
9	1	2	6
10	1	2	8
11	2	1	10
12	2	1	8
13	2	1	7
14	2	1	7
15	2	1	9
16	2	2	7
17	2	2	9
18	2	2	12
19	2	2	8
20	2	2	8
21	3	1	7
22	3	1	5
23	3	1	6
24	3	1	5
25	3	1	9
26	3	2	9
27	3	2	7
28	3	2	5
29	3	2	4
30	3	2	6
31	4	1	8
32	4	1	4
33	4	1	6
34	4	1	5
35	4	1	5
36	4	2	5
37	4	2	7
38	4	2	9
39	4	2	7
40	4	2	10

Con el comando `MTB < Twoway 'Defects' 'Machine' 'Shifts'` se obtiene el análisis de varianza en dos sentidos. El valor p para la interacción es 0.494. Éste es el nivel de significancia mínimo para rechazar la hipótesis nula; es claro que no hay una interacción significativa entre turnos y máquinas. Para los turnos, el valor p es 0.090; como este valor es mayor que 0.050, la cantidad media de defectos en los dos turnos no es significativamente diferente. Para las máquinas,

el valor p es 0.002; al nivel de significancia 0.050, las cantidades medias de defectos en las cuatro máquinas son notablemente diferentes.

```
MTB > Twoway 'Defects' 'Machine' 'Shift'
```

Two-way Analysis of Variance

```
Analysis of Variance for Defects
Source          DF        SS        MS        F        P
Machine          3     51.00     17.00     6.42    0.002
Shift            1      8.10      8.10     3.06    0.090
Interaction      3      6.50      2.17     0.82    0.494
Error           32     84.80      2.65
Total           39    150.40
```

En la figura 16-7 se presenta la gráfica de la interacción entre turnos y máquinas. La gráfica indica que hay una posible interacción entre turnos y máquinas. Sin embargo, en la tabla del análisis de varianza el valor p de esta interacción indica que no hay una interacción significativa. Cuando no hay interacción, las gráficas del turno 1 y del turno 2 son paralelas. En la figura 16-8, las gráficas de los efectos principales indican que, en este experimento, la máquina 1 fue la que produjo en promedio menos tornillos defectuosos y la máquina 2 fue la que produjo más tornillos defectuosos; se produjeron más defectuosos en el turno 2 que en el turno 1. Sin embargo, el análisis de varianza indica que esta diferencia no es significativa.

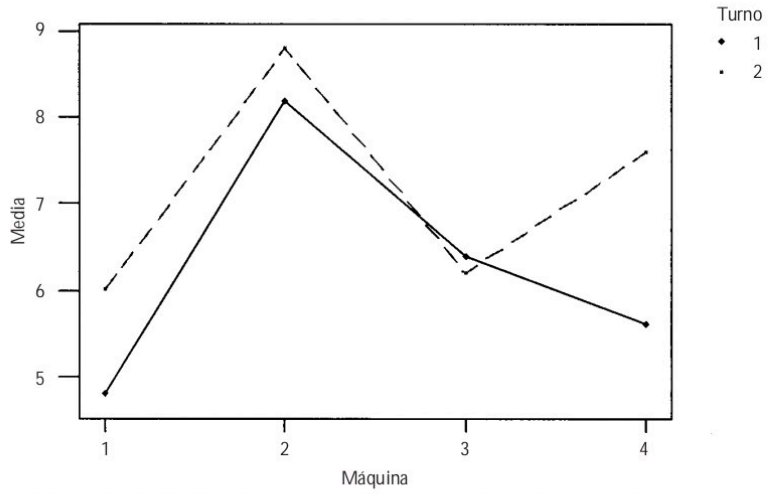


Figura 16-7 Gráfica de la interacción: medias de los datos para defectuosos.

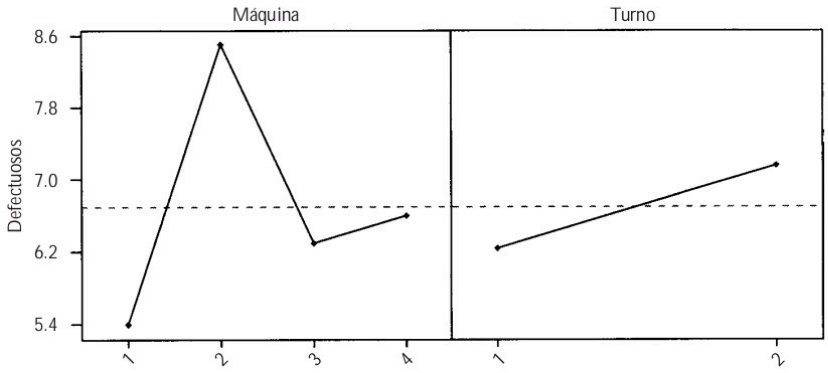


Figura 16-8 Gráfica de los efectos principales: medias de los datos para defectuosos.

16.14 Resolver el problema 16.13 empleando EXCEL.

SOLUCIÓN

A	B	C	D	E
	máquina1	máquina2	máquina3	máquina4
turno1	6	10	7	8
turno1	4	8	5	4
turno1	5	7	6	6
turno1	5	7	5	5
turno1	4	9	9	5
turno2	5	7	9	5
turno2	7	9	7	7
turno2	4	12	5	9
turno2	6	8	4	7
turno2	8	8	6	10

Análisis de varianza de dos factores con una sola muestra por grupo

RESUMEN	máquina1	máquina2	máquina3	máquina4	Total
turno1					
Cuenta	5	5	5	5	20
Suma	24	41	32	28	125
Promedio	4.8	8.2	6.4	5.6	6.25
Varianza	0.7	1.7	2.8	2.3	3.25
turno2					
Cuenta	5	5	5	5	20
Suma	30	44	31	38	143
Promedio	6	8.8	6.2	7.6	7.15
Varianza	2.5	3.7	3.7	3.8	4.239474
Total					
Cuenta	10	10	10	10	
Suma	54	85	63	66	
Promedio	5.4	8.5	6.3	6.6	
Varianza	1.822222	2.5	2.9	3.822222	

ANÁLISIS DE VARIANZA

Origen de las variaciones	*SS*	*df*	*MS*	*F*	*Valor p*
Muestra	8.1	1	8.1	3.056604	0.089999
Columnas	51	3	17	6.415094	0.001584
Interacción	6.5	3	2.166667	0.81761	0.49371
Dentro del grupo	84.8	32	2.65		
Total	150.4	39			

**SS*, Suma de cuadrados; *df*, grados de libertad; *MS*, promedio de los cuadrados; *F*, probabilidad.

Los datos se ingresan en la hoja de cálculo de EXCEL, como se muestra. Con la secuencia **Tools → Data analysis → Anova: Two-Factor with Replication** se obtiene el cuadro de diálogo de la figura 16-9 que se llena como se muestra en la figura.

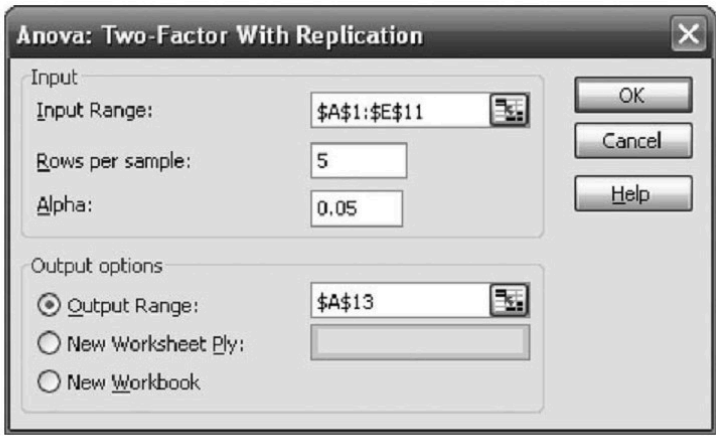


Figura 16-9 EXCEL, cuadro de diálogo para el problema 16.14.

En el análisis de varianza, muestra corresponde a turnos, y columna a máquinas. Comparar estos resultados con los obtenidos con MINITAB en el problema 16.13.

CUADRADOS LATINOS

16.15 Un granjero desea probar los efectos de cuatro fertilizantes (*A*, *B*, *C* y *D*) en la producción de trigo. Con objeto de eliminar las fuentes de error debidas a la variabilidad de la fertilidad del suelo, distribuye los fertilizantes en un cuadrado latino, como se muestra en la tabla 16.24, en donde los números indican la producción en bushels por unidad de área. Hacer un análisis de varianza para determinar, a los niveles de significancia *a*) 0.05 y *b*) 0.01, si hay diferencia entre los fertilizantes. *c*) Proporcionar la solución de MINITAB para este diseño de cuadrado latino. *d*) Proporcionar la solución de STATISTIX para este diseño de cuadrado latino.

SOLUCIÓN

Primero, como se muestra en la tabla 16.25, se obtienen los totales de los renglones y los totales de las columnas. También se obtiene la producción total obtenida con cada uno de los fertilizantes, como se muestra en la tabla 16.26. Después, como es costumbre, se obtienen la variación total y las variaciones de los renglones, de las columnas y de los tratamientos. Se encuentra que:

La variación total es

$$V = (18)^2 + (21)^2 + (25)^5 + \cdots + (10)^2 + (17) - \frac{(295)^2}{16}$$

$$= 5\,769 - 5\,439.06 = 329.94$$

Tabla 16.24

A 18	*C* 21	*D* 25	*B* 11
D 22	*B* 12	*A* 15	*C* 19
B 15	*A* 20	*C* 23	*D* 24
C 22	*D* 21	*B* 10	*A* 17

Tabla 16.25

					Total
	A 18	*C* 21	*D* 25	*B* 11	75
	D 22	*B* 12	*A* 15	*C* 19	68
	B 15	*A* 20	*C* 23	*D* 24	82
	C 22	*D* 21	*B* 10	*A* 17	70
Total	77	74	73	71	295

Tabla 16.26

	A	*B*	*C*	*D*	
Total	70	48	85	92	295

La variación entre los renglones es

$$V_R = \frac{(75)^2}{4} + \frac{(68)^2}{4} + \frac{(82)^2}{4} + \frac{(70)^2}{4} - \frac{(295)^2}{16}$$
$$= 5\,468.25 - 5\,439.06 = 29.19$$

La variación entre las columnas es

$$V_C = \frac{(77)^2}{4} + \frac{(74)^2}{4} + \frac{(73)^2}{4} + \frac{(71)^2}{4} - \frac{(295)^2}{16}$$
$$= 5\,443.75 - 5\,439.06 = 4.69$$

La variación entre los tratamientos es

$$V_B = \frac{(70)^2}{4} + \frac{(48)^2}{4} + \frac{(85)^2}{4} + \frac{(92)^2}{4} - \frac{(295)^2}{16} = 5\,723.25 - 5\,439.06 = 284.19$$

En la tabla 16.27 se muestra el análisis de varianza.

Tabla 16.27

Variación	Grados de libertad	Cuadrado medio	*F*
Renglones, 29.19	3	9.73	4.92
Columnas, 4.69	3	1.563	0.79
Tratamientos, 284.19	3	94.73	47.9
Residuales, 11.87	6	1.978	
Total, 329.94	15		

a) Como $F_{.95,3,6} = 4.76$, la hipótesis de que las medias de los renglones son iguales puede rechazarse al nivel de significancia 0.05. Al nivel 0.05 existe diferencia en la fertilidad del suelo entre un renglón y otro.

Como el valor F para las columnas es menor a 1, se concluye que en las columnas no hay diferencia en la fertilidad del suelo.

Como el valor F para los tratamientos es $47.9 > 4.76$, se concluye que hay diferencia entre los fertilizantes.

b) Como $F_{.99,3,6} = 9.78$, al nivel de significancia 0.01 se puede aceptar la hipótesis de que no hay diferencia en la fertilidad del suelo en los renglones (o en las columnas). Sin embargo, al nivel de significancia 0.01 se sigue concluyendo que hay diferencia entre los fertilizantes.

c) Primero se presenta la estructura que deben tener los datos en la hoja de cálculo de MINITAB.

Row	Rows	Columns	Treatment	Yield
1	1	1	1	18
2	1	2	3	21
3	1	3	4	25
4	1	4	2	11
5	2	1	4	22
6	2	2	2	12
7	2	3	1	15
8	2	4	3	19
9	3	1	2	15
10	3	2	1	20
11	3	3	3	23
12	3	4	4	24
13	4	1	3	22
14	4	2	4	21
15	4	3	2	10
16	4	4	1	17

Obsérvese que los renglones y las columnas se han numerado del 1 al 4. Los fertilizantes *A* a *D* de la tabla 16.24 han sido codificados en la hoja de cálculo 1 al 4, respectivamente. Con la secuencia de MINITAB **Stat** → **ANOVA** → **General Linear Model** se obtiene el resultado siguiente.

Modelo lineal general: rendimiento *versus* filas, columnas, tratamiento

Factor	Type	Levels	Values
Rows	fixed	4	1, 2, 3, 4
Columns	fixed	4	1, 2, 3, 4
Treatment	fixed	4	1, 2, 3, 4

Analysis of Variance for Yield, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Rows	3	29.188	29.188	9.729	4.92	0.047
Columns	3	4.688	4.687	1.562	0.79	0.542
Treatment	3	284.188	284.188	94.729	47.86	0.000
Error	6	11.875	11.875	1.979		
Total	15	329.938				

S=1.40683 R-Sq=96.40% R-sq(adj)-91.00%

Los resultados de MINITAB coinciden con los antes obtenidos a mano. Estos resultados indican que hay diferencia en la fertilidad de un renglón a otro, al nivel de significancia 0.05, pero no al nivel 0.01. No hay diferencia en la fertilidad de una columna a otra. Hay diferencia entre los cuatro fertilizantes al nivel de significancia 0.01.

d) Con la secuencia **Statistics** → **Linear Models** → **Analysis of Variance** → **Latin Square Design** de STATISTIX se obtiene el resultado siguiente.

Statistix 8.0

Latin Square AOV Table for Yield

Source	DF	SS	MS	F	P
Rows	3	29.188	9.7292		
Columns	3	4.688	1.5625		
Treatment	3	284.188	94.7292	47.86	0.0001
Error	6	11.875	1.9792		
Total	15	329.938			

CUADRADOS GRECOLATINOS

16.16 Se quiere determinar si hay una diferencia significativa entre las gasolinas *A*, *B*, *C* y *D* en su rendimiento por galón. Diseñar un experimento en el que se usen cuatro conductores distintos, cuatro automóviles distintos y cuatro carreteras distintas.

SOLUCIÓN

Como el número de gasolinas, de conductores, de automóviles y de carreteras es el mismo (cuatro), puede emplearse un cuadrado grecolatino. Supóngase que los diferentes automóviles están representados por los renglones y los diferentes conductores por las columnas, como se muestra en la tabla 16.28. Después, las gasolinas (*A*, *B*, *C* y *D*) se asignan en forma aleatoria a los renglones y a las columnas, sujetas a la condición de que cada letra aparezca sólo una vez en cada renglón y una vez en cada columna. Por lo tanto, cada conductor tendrá la oportunidad de conducir cada uno de los automóviles y usar cada uno de los tipos de gasolina, y ningún automóvil será conducido dos veces con el mismo tipo de gasolina.

Ahora se asignan en forma aleatoria las cuatro carreteras, denotadas α, β, γ y δ, sujetándolas a la misma condición impuesta a los cuadrados latinos. Por lo tanto, cada conductor tendrá también la oportunidad de conducir por cada una de las carreteras. En la tabla 16.28 se muestra una posible distribución.

Tabla 16.28

	Conductor			
	1	2	3	4
Automóvil 1	B_γ	A_β	D_δ	C_α
Automóvil 2	A_δ	B_α	C_γ	D_β
Automóvil 3	D_α	C_δ	B_β	A_γ
Automóvil 4	C_β	D_γ	A_α	B_δ

16.17 Supóngase que al llevar a cabo el experimento del problema 16.16, las millas por galón son las indicadas en la tabla 16.29. Utilizar el análisis de varianza para determinar si al nivel de significancia 0.05 hay diferencias. Usar MINITAB para obtener la tabla del análisis de varianza y usar los valores *p* dados por MINITAB para probar si existen diferencias al nivel de significancia 0.05.

Tabla 16.29

	Conductor			
	1	2	3	4
Automóvil 1	B_γ 19	A_β 16	D_δ 16	C_α 14
Automóvil 2	A_δ 15	B_α 18	C_γ 11	D_β 15
Automóvil 3	D_α 14	C_δ 11	B_β 21	A_γ 16
Automóvil 4	C_β 16	D_γ 16	A_α 15	B_δ 23

SOLUCIÓN

Primero se obtienen los totales de los renglones y de las columnas, como se muestra en la tabla 16.30. Después se obtienen los totales correspondientes a cada letra latina y a cada letra griega como se indica a continuación:

$$\text{Total } A: \quad 15 + 16 + 15 + 16 = 62$$

$$\text{Total } B: \quad 19 + 18 + 21 + 23 = 81$$

$$\text{Total } C: \quad 16 + 11 + 11 + 14 = 52$$

$$\text{Total } D: \quad 14 + 16 + 16 + 15 = 61$$

$$\text{Total } \alpha: \quad 14 + 18 + 15 + 14 = 61$$

$$\text{Total } \beta: \quad 16 + 16 + 21 + 15 = 68$$

$$\text{Total } \gamma: \quad 19 + 16 + 11 + 16 = 62$$

$$\text{Total } \delta: \quad 15 + 11 + 16 + 23 = 65$$

Ahora, se calculan las variaciones correspondientes, usando el método abreviado:

$$\text{Renglones:} \quad \frac{(65)^2}{4} + \frac{(59)^2}{4} + \frac{(62)^2}{4} + \frac{(70)^2}{4} - \frac{(256)^2}{16} = 4\,112.50 - 4\,096 = 16.50$$

$$\text{Columnas:} \quad \frac{(64)^2}{4} + \frac{(61)^2}{4} + \frac{(63)^2}{4} + \frac{(68)^2}{4} - \frac{(256)^2}{16} = 4\,102.50 - 4\,096 = 6.50$$

$$\text{Gasolinas } (A, B, C, D): \quad \frac{(62)^2}{4} + \frac{(81)^2}{4} + \frac{(52)^2}{4} + \frac{(61)^2}{4} - \frac{(256)^2}{16} = 4\,207.50 - 4\,096 = 111.50$$

$$\text{Carreteras } (\alpha, \beta, \gamma, \delta): \quad \frac{(61)^2}{4} + \frac{(68)^2}{4} + \frac{(62)^2}{4} + \frac{(65)^2}{4} - \frac{(256)^2}{16} = 4\,103.50 - 4\,096 = 7.50$$

La variación total es

$$(19)^2 + (16)^2 + (16)^2 + \cdots + (15)^2 + (23)^2 - \frac{(256)^2}{16} = 4\,244 - 4\,096 = 148.00$$

de manera que la variación debida al error es

$$148.00 - 16.50 - 6.50 - 111.50 - 7.50 = 6.00$$

Los resultados se muestran en la tabla 16.31, la tabla del análisis de varianza. El número total de grados de libertad es $N^2 - 1$, ya que se trata de un cuadrado de $N \times N$. Cada uno de los renglones, de las columnas, de las letras latinas y de las letras griegas tiene $N - 1$ grados de libertad. Por lo tanto, los grados de libertad para el error son $N^2 - 1 - 4(N - 1) = (N - 1)(N - 3)$. En este caso, $N = 4$.

Tabla 16.30

					Total
	B_γ 19	A_β 16	D_δ 16	C_α 14	65
	A_δ 15	B_α 18	C_γ 11	D_β 15	59
	D_α 14	C_δ 11	B_β 21	A_γ 16	62
	C_β 16	D_γ 16	A_α 15	B_δ 23	70
Total	64	61	63	68	256

Tabla 16.31

Variación	Grados de libertad	Cuadrado medio	F
Renglones (automóviles), 16.50	3	5.500	$\frac{5.500}{2.000} = 2.75$
Columnas (conductores), 6.50	3	2.167	$\frac{2.167}{2.000} = 1.08$
Gasolinas (A, B, C y D), 111.50	3	37.167	$\frac{37.167}{2.000} = 18.6$
Carreteras (α, β, γ y δ), 7.50	3	2.500	$\frac{2.500}{2.000} = 1.25$
Error, 6.00	3	2.000	
Total, 148.00	15		

Se tiene que $F_{.95,3,3} = 9.28$ y $F_{.99,3,3} = 29.5$. Por lo tanto, la hipótesis de que las gasolinas son iguales puede rechazarse al nivel 0.05, pero no al nivel 0.01.

Primero se presenta la estructura que deben tener los datos en la hoja de cálculo de MINITAB.

Row	Car	Driver	Gasoline	Road	MPG
1	1	1	2	3	19
2	1	2	1	2	16
3	1	3	4	4	16
4	1	4	3	1	14
5	2	1	1	4	15
6	2	2	2	1	18
7	2	3	3	3	11
8	2	4	4	2	15
9	3	1	4	1	14
10	3	2	3	4	11
11	3	3	2	2	21
12	3	4	1	3	16
13	4	1	3	2	16
14	4	2	4	3	16
15	4	3	1	1	15
16	4	4	2	4	23

Obsérvese que los automóviles y los conductores están numerados en la hoja de cálculo de MINITAB igual que en la tabla 16.29. Las gasolinas en la tabla 16.29 van de la A a la D, y en la hoja de cálculo de MINITAB se han codificado del 1 al 4, respectivamente. Las carreteras son α, β, γ y δ en la tabla 16.29, en la hoja de cálculo de MINITAB se han codificado 1, 2, 3 y 4. Con la secuencia **Stat** → **ANOVA** → **General Linear Model** de MINITAB se obtienen los resultados siguientes.

Modelo lineal general: millas por galón *versus* automóvil, conductor, gasolina, carretera

Factor	Type	Levels	Values
Car	fixed	4	1, 2, 3, 4
Driver	fixed	4	1, 2, 3, 4
Gasoline	fixed	4	1, 2, 3, 4
Road	fixed	4	1, 2, 3, 4

```
Analysis of Variance for Mpg, using Adjusted SS for Tests

Source      DF      Seq SS    Adj SS    Adj MS        F        P
Car          3      16.500    16.500     5.500     2.75    0.214
Driver       3       6.500     6.500     2.167     1.08    0.475
Gasoline     3     111.500   111.500    37.167    18.58    0.019
Road         3       7.500     7.500     2.500     1.25    0.429
Error        3       6.000     6.000     2.000
Total       15     148.000
```

La columna titulada `Seq MS` en los resultados de MINITAB corresponde a la columna titulada `Mean Square` en la tabla 16.31. Los valores *F* `calculados` en los resultados de MINITAB son los mismos que los de la tabla 16.31. Los valores *p* para automóviles, conductores, marcas de gasolina y carreteras son 0.214, 0.475, 0.019 y 0.429, respectivamente. Recuérdese que un valor *p* es el mínimo valor para un nivel de significancia preestablecido al que puede rechazarse la hipótesis de medias iguales de un factor. Los valores *p* indican que no hay diferencia entre los automóviles, conductores o carreteras a los niveles 0.01 o 0.05. Las medias de las marcas de gasolina son estadísticamente diferentes al nivel 0.05, pero no al nivel 0.01. Posteriores investigaciones sobre las medias de las marcas de gasolina pueden indicar cómo éstas difieren.

PROBLEMAS DIVERSOS

16.18 Probar [ecuación (*15*) de este capítulo] que $\sum_j \alpha_j = 0$.

SOLUCIÓN

Las medias de la población de los tratamientos μ_j y la media de la población total μ se relacionan mediante

$$\mu = \frac{1}{a}\sum_j \mu_j \tag{53}$$

Entonces, como $\alpha_j = \mu_j = \mu$, empleando la ecuación (*53*) se tiene,

$$\sum_j \alpha_j = \sum_j (\mu_j - \mu) = \sum_j \mu_j - a\mu = 0 \tag{54}$$

16.19 Deducir: *a*) la ecuación (*16*) y *b*) la ecuación (*17*) de este capítulo.

SOLUCIÓN

a) Por definición, se tiene

$$V_W = \sum_{j,k}(X_{jk} - \bar{X}_{j.})^2 = b\sum_{j=1}^{a}\left[\frac{1}{b}\sum_{k=1}^{b}(X_{jk} - \bar{X}_{j.})^2\right] = b\sum_{j=1}^{a} S_j^2$$

donde S_j^2 es la varianza muestral correspondiente al tratamiento *j*. Entonces, como el tamaño de la muestra es *b*,

$$E(V_W) = b\sum_{j=1}^{a} E(S_j^2) = b\sum_{j=1}^{a}\left(\frac{b-1}{b}\sigma^2\right) = a(b-1)\sigma^2$$

b) Por definición

$$V_B = b\sum_{j=1}^{a}(\bar{X}_{j.} - \bar{X})^2 = b\sum_{j=1}^{a}\bar{X}_{j.}^2 - 2b\bar{X}\sum_{j=1}^{a}\bar{X}_{j.} + ab\bar{X}^2 = b\sum_{j=1}^{a}\bar{X}_{j.}^2 - ab\bar{X}^2$$

ya que $\bar{X} = (\sum_j \bar{X}_{j.})/a$. Entonces, omitiendo el índice de sumación se tiene

$$E(V_B) = b\sum E(\bar{X}_{j.}^2) - abE(\bar{X}^2) \tag{55}$$

Ahora para cualquier variable aleatoria U, $E(U^2) = \text{var}(U) + [E(U)]^2$, donde var($U$) denota la varianza de U. Por lo tanto,

$$E(\bar{X}_{j.}^2) = \text{var}\,(\bar{X}_{j.}) + [E(\bar{X}_{j.})]^2 \tag{56}$$

$$E(\bar{X}^2) = \text{var}\,(\bar{X}) + [E(\bar{X})]^2 \tag{57}$$

Pero como las poblaciones de los tratamientos son normales, con media $\mu_j = \mu + \alpha_j$, se tiene

$$\text{var}\,(\bar{X}_{j.}) = \frac{\sigma^2}{b} \tag{58}$$

$$\text{var}\,(\bar{X}) = \frac{\sigma^2}{ab} \tag{59}$$

$$E(\bar{X}_{j.}) = \mu_j = \mu + \alpha_j \tag{60}$$

$$E(\bar{X}) = \mu \tag{61}$$

Usando las ecuaciones (*56*) a (*61*) junto con la ecuación (*55*), se tiene

$$\begin{aligned} E(V_B) &= b\sum\left[\frac{\sigma^2}{b} + (\mu + \alpha_j)^2\right] - ab\left[\frac{\sigma^2}{ab} + \mu^2\right] \\ &= a\sigma^2 + b\sum(\mu + \alpha_j)^2 - \sigma^2 - ab\mu^2 \\ &= (a-1)\sigma^2 + ab\mu^2 + 2b\mu\sum\alpha_j + b\sum\alpha_j^2 + ab\mu^2 \\ &= (a-1)\sigma^2 + b\sum\alpha_j^2 \end{aligned}$$

16.20 Probar el teorema 1 de este capítulo.

SOLUCIÓN

Como se muestra en el problema 16.19,

$$V_W = b\sum_{j=1}^{a} S_j^2 \qquad \text{o bien} \qquad \frac{V_W}{\sigma^2} = \sum_{j=1}^{a}\frac{bS_j^2}{\sigma^2}$$

donde S_j^2 es la varianza muestral de muestras de tamaño b obtenidas de la población del tratamiento j. Se ve que bS_j^2/σ^2 tiene una distribución ji cuadrada con $b-1$ grados de libertad. Por lo tanto, como las varianzas S_j^2 son independientes, se concluye, de acuerdo con la página 299, que V_W/σ^2 tiene una distribución ji cuadrada con $a(b-1)$ grados de libertad.

PROBLEMAS SUPLEMENTARIOS

CLASIFICACIÓN EN UN SENTIDO O EXPERIMENTOS CON UN FACTOR

Se aconseja al lector que todos estos ejercicios los haga primero "a mano", empleando las ecuaciones dadas en este capítulo, antes de usar el software sugerido. Esto le ayudará a comprender mejor la técnica ANOVA, así como a apreciar la potencia del software.

16.21 Se realiza un experimento para determinar el rendimiento de cinco tipos diferentes de trigo: *A*, *B*, *C*, *D* y *E*. A cada variedad se le asignan cuatro parcelas; los rendimientos (en bushels por acre) se muestran en la tabla 16.32. Suponiendo que las parcelas tengan una fertilidad semejante y que las variedades de trigo se asignen a las parcelas en forma aleatoria, a los niveles de significancia *a*) 0.05 y *b*) 0.01, determinar si hay diferencia entre los rendimientos. *c*) Proporcionar el análisis que se obtiene para esta clasificación en un sentido o experimento de un factor empleando MINITAB.

Tabla 16.32

A	20	12	15	19
B	17	14	12	15
C	23	16	18	14
D	15	17	20	12
E	21	14	17	18

16.22 Una empresa quiere probar cuatro tipos de neumáticos: *A*, *B*, *C* y *D*. En la tabla 16.33 se da (en miles de millas) la duración de estos neumáticos, determinada por el dibujo, donde cada tipo de neumático ha sido probado en seis automóviles similares asignados, a los neumáticos, en forma aleatoria. A los niveles de significancia: *a*) 0.05 y *b*) 0.01, determinar si hay alguna diferencia significativa entre los neumáticos. *c*) Proporcionar el análisis que se obtiene para esta clasificación en un sentido o experimento de un factor empleando STATISTIX.

Tabla 16.33

A	33	38	36	40	31	35
B	32	40	42	38	30	34
C	31	37	35	33	34	30
D	29	34	32	30	33	31

16.23 Un maestro quiere probar tres métodos de enseñanza: I, II y III. Para esto se eligen en forma aleatoria tres grupos, cada uno de cinco estudiantes y con cada grupo se emplea uno de estos tres métodos de enseñanza. A todos los estudiantes se les aplica un mismo examen. En la tabla 16.34 se presentan las calificaciones que obtuvieron. A los niveles de significancia: *a*) 0.05 y *b*) 0.01, determinar si hay diferencia entre estos tres métodos de enseñanza. *c*) Proporcionar el análisis que se obtiene empleando EXCEL para esta clasificación en un sentido o experimento de un factor.

Tabla 16.34

Método I	75	62	71	58	73
Método II	81	85	68	92	90
Método III	73	79	60	75	81

MODIFICACIONES PARA NÚMEROS DISTINTOS DE OBSERVACIONES

16.24 En la tabla 16.35 se dan las cifras en millas por galón obtenidas en automóviles similares usando cinco marcas de gasolina. A los niveles de significancia: *a*) 0.05 y *b*) 0.01, determinar si hay diferencia entre las marcas. *c*) Proporcionar el análisis que se obtiene para esta clasificación en un sentido o experimento de un factor empleando SPSS.

Tabla 16.35

Marca *A*	12	15	14	11	15
Marca *B*	14	12	15		
Marca *C*	11	12	10	14	
Marca *D*	15	18	16	17	14
Marca *E*	10	12	14	12	

Tabla 16.36

Matemáticas	72	80	83	75	
Ciencias	81	74	77		
Inglés	88	82	90	87	80
Economía	74	71	77	70	

16.25 En la tabla 16.36 se presentan las calificaciones obtenidas durante un semestre por un estudiante. A los niveles de significancia: *a*) 0.05 y *b*) 0.01, determinar si hay diferencia significativa entre las calificaciones de este estudiante. *c*) Proporcionar el análisis que se obtiene para esta clasificación en un sentido o experimento de un factor empleando SAS.

CLASIFICACIÓN EN DOS SENTIDOS O EXPERIMENTOS CON DOS FACTORES

16.26 Los artículos que produce una empresa son fabricados por tres operadores que usan tres máquinas diferentes. La empresa desea determinar si hay diferencia: *a*) entre los operadores y *b*) entre las máquinas. Se realiza un experimento para determinar la cantidad de artículos por día producidos por cada operador usando cada máquina; en la tabla 16.37 se muestran los resultados. Empleando MINITAB al nivel de significancia 0.05, proporcionar la información buscada.

Tabla 16.37

	Operador		
	1	2	3
Máquina *A*	23	27	24
Máquina *B*	34	30	28
Máquina *C*	28	25	27

Tabla 16.38

	Tipo de trigo			
	I	II	III	IV
Bloque *A*	12	15	10	14
Bloque *B*	15	19	12	11
Bloque *C*	14	18	15	12
Bloque *D*	11	16	12	16
Bloque *E*	16	17	11	14

16.27 Resolver el problema 16.26 usando EXCEL y el nivel de significancia 0.01.

16.28 Se plantan semillas de cuatro tipos diferentes de trigo en cinco bloques. Cada bloque se divide en cuatro parcelas que después se asignan en forma aleatoria a los cuatro tipos de trigo. Al nivel de significancia 0.05, determinar si los rendimientos, en bushels por acre, que se presentan en la tabla 16.38, varían en forma significativa con respecto a: *a*) el suelo (es decir, a los cinco bloques) y *b*) el tipo de trigo. Usar SPSS para construir una tabla de ANOVA.

16.29 Resolver el problema 16.28 usando STATISTIX y el nivel de significancia 0.01 para construir ANOVA.

16.30 Supóngase que en el problema 16.22 la primera observación con cada tipo de neumático se haga usando determinado tipo de automóvil, la segunda observación se haga usando un segundo tipo de automóvil, y así sucesivamente. Determinar al

nivel de significancia 0.05 si hay diferencia: *a*) entre los tipos de neumáticos y *b*) entre los tipos de automóviles. Usar SAS para construir la tabla de ANOVA.

16.31 Resolver el problema 16.30 usando MINITAB y el nivel de significancia 0.01.

16.32 Supóngase que en el problema 16.23 la primera entrada para cada método de enseñanza corresponde a un estudiante de determinada escuela, la segunda a un estudiante de otra escuela, y así sucesivamente. Probar la hipótesis de que al nivel de significancia 0.05 hay diferencia: *a*) entre los métodos de enseñanza y *b*) entre las escuelas. Para construir una tabla de ANOVA usar STATISTIX.

16.33 Se realiza un experimento para probar si el color de pelo y la estatura de las estudiantes de Estados Unidos tiene alguna relación con los logros escolares. En la tabla 16.39 se presentan los resultados, donde los números indican la cantidad de personas en el 10% superior de esta evaluación. Analizar el experimento al nivel de significancia 0.05. Para construir una tabla de ANOVA utilizar EXCEL.

Tabla 16.39

	Pelirroja	Rubia	Castaña
Alta	75	78	80
Mediana	81	76	79
Baja	73	75	77

Tabla 16.40

A	16	18	20	23
B	15	17	16	19
C	21	19	18	21
D	18	22	21	23
E	17	18	24	20

16.34 Resolver el problema 16.33 al nivel de significancia 0.01. Usar SPSS para obtener la tabla de ANOVA y comparar los resultados con los de EXCEL dados en el problema 16.33.

EXPERIMENTOS CON DOS FACTORES CON REPLICACIÓN

16.35 Supóngase que el experimento del problema 16.21 se llevó a cabo en el sur y que las columnas de la tabla 16.32 indican ahora cuatro tipos de fertilizantes, y que un experimento similar se lleva a cabo en el oeste dando los resultados que se muestran en la tabla 16.40. Al nivel de significancia 0.05, determinar si hay diferencia en los rendimientos que se deban a: *a*) los fertilizantes y *b*) el lugar. Usar MINITAB para elaborar la tabla de ANOVA.

16.36 Resolver el problema 16.35 al nivel de significancia 0.01. Para elaborar la tabla de ANOVA emplear STATISTIX y comparar los resultados con los dados por MINITAB en el problema 16.35.

16.37 En la tabla 16.41 se dan las cantidades de artículos producidos, en cada uno de los días de la semana, por cuatro operadores que trabajan con dos tipos de máquinas, I y II. Al nivel de significancia 0.05, determinar si hay diferencias significativas: *a*) entre los operadores y *b*) entre las máquinas. Construir una tabla de ANOVA usando SAS y otra usando MINITAB.

Tabla 16.41

	Máquina I					Máquina II				
	Lunes	Martes	Miércoles	Jueves	Viernes	Lunes	Martes	Miércoles	Jueves	Viernes
Operador *A*	15	18	17	20	12	14	16	18	17	15
Operador *B*	12	16	14	18	11	11	15	12	16	12
Operador *C*	14	17	18	16	13	12	14	16	14	11
Operador *D*	19	16	21	23	18	17	15	18	20	17

CUADRADOS LATINOS

16.38 Se realiza un experimento para probar los efectos sobre la producción de trigo de cuatro fertilizantes (*A*, *B*, *C* y *D*) y de las variaciones en el suelo en dos direcciones perpendiculares. Se obtiene el cuadrado latino de la tabla 16.42, donde los números corresponden a la producción de trigo por unidad de área. Al nivel de significancia 0.01, probar la hipótesis de que no hay diferencia entre: *a*) los fertilizantes y *b*) las variaciones en el suelo. Usar STATISTIX para elaborar la tabla de ANOVA.

Tabla 16.42

C 8	*A* 10	*D* 12	*B* 11
A 14	*C* 12	*B* 11	*D* 15
D 10	*B* 14	*C* 16	*A* 10
B 7	*D* 16	*A* 14	*C* 12

Tabla 16.43

E 75	*W* 78	*M* 80
M 81	*E* 76	*W* 79
W 73	*M* 75	*E* 77

16.39 Resolver el problema 16.38 al nivel de significancia 0.05. Para construir la tabla ANOVA utilizar MINITAB y comparar los resultados con los de STATISTIX del problema 16.38.

16.40 Volviendo al problema 16.33, suponer que se introduce un factor más, dado por el área *E*, *M* o *W* de Estados Unidos en la que nació la estudiante, como se muestran en la tabla 16.43. Al nivel de significancia 0.05, determinar si hay diferencia significativa en los logros académicos de las estudiantes debido a: *a*) la estatura, *b*) el color de pelo y *c*) el lugar de nacimiento. Usar SPSS para elaborar la tabla de ANOVA.

CUADRADOS GRECOLATINOS

16.41 Con objeto de producir un tipo mejor de alimento para gallinas, a los ingredientes básicos se les agregan cuatro cantidades distintas de cada una de dos sustancias químicas. Las diferentes cantidades de la primera sustancia química se indican como *A*, *B*, *C* y *D*, en tanto que las cantidades de la segunda sustancia química se indican como α, β, γ y δ. El alimento es suministrado a pollitos recién nacidos agrupados de acuerdo con cuatro pesos iniciales (W_1, W_2, W_3 y W_4) y a cuatro especies diferentes (S_1, S_2, S_3 y S_4). En el cuadro grecolatino de la tabla 16.44 se da el aumento de peso por unidad de tiempo. Efectuar un análisis de varianza de este experimento al nivel de significancia 0.05 y brindar las conclusiones que se obtengan. Para elaborar la tabla de ANOVA, usar MINITAB.

Tabla 16.44

	W_1	W_2	W_3	W_4
S_1	C_γ 8	B_β 6	A_α 5	D_δ 6
S_2	A_δ 4	D_α 3	C_β 7	B_γ 3
S_3	D_β 5	A_γ 6	B_δ 5	C_α 6
S_4	B_α 6	C_δ 10	D_γ 10	A_β 8

16.42 Cada una de cuatro empresas (C_1, C_2, C_3 y C_4) fabrica cuatro tipos distintos de cable (T_1, T_2, T_3 y T_4). Cuatro operadores (*A*, *B*, *C* y *D*) emplean cuatro máquinas diferentes (α, β, γ y δ) para medir la resistencia de los cables. Las resistencias

promedio encontradas se presentan en el cuadro grecolatino de la tabla 16.45. Al nivel de significancia 0.05, hacer un análisis de varianza y proporcionar las conclusiones. Usar SPSS para elaborar la tabla de ANOVA.

PROBLEMAS DIVERSOS

16.43 En la tabla 16.46 se dan los datos del óxido acumulado sobre hierro tratado con las sustancias químicas *A*, *B* y *C*, respectivamente. A los niveles de significancia: *a*) 0.05 y *b*) 0.01, determinar si hay diferencia significativa entre los tratamientos. Usar EXCEL para elaborar la tabla de ANOVA.

Tabla 16.45

	C_1	C_2	C_3	C_4
T_1	A_β 164	B_γ 181	C_α 193	D_δ 160
T_2	C_δ 171	D_α 162	A_γ 183	B_β 145
T_3	D_γ 198	C_β 212	B_δ 207	A_α 188
T_4	B_α 157	A_δ 172	D_β 166	C_γ 136

Tabla 16.46

A	3	5	4	4
B	4	2	3	3
C	6	4	5	5

16.44 En un experimento se mide el coeficiente intelectual (CI) de estudiantes adultos de estaturas baja, media y alta. En la tabla 16.47 se dan los resultados. A los niveles de significancia: *a*) 0.05 y *b*) 0.01, determinar si hay diferencia en los CI de acuerdo con las distintas estaturas. Usar MINITAB para elaborar la tabla de ANOVA.

Tabla 16.47

Alta	110	105	118	112	90	
Baja	95	103	115	107		
Media	108	112	93	104	96	102

16.45 Probar las ecuaciones (*10*), (*11*) y (*12*) de este capítulo.

16.46 Se hace un examen para determinar, entre veteranos y no veteranos con diferente CI, quiénes tienen mejor desempeño. Las puntuaciones obtenidas se muestran en la tabla 16.48. Al nivel de significancia 0.05, determinar si hay diferencias en las puntuaciones debidas a diferencias en: *a*) ser o no veterano y *b*) el CI. Usar SPSS para elaborar una tabla de ANOVA.

Tabla 16.48

	Puntuación en la prueba		
	CI alto	CI medio	CI bajo
Veterano	90	81	74
No veterano	85	78	70

16.47 Usar STATISTIX para resolver el problema 16.46 al nivel de significancia 0.01.

16.48 En la tabla 16.49 se presentan las puntuaciones que obtuvieron en un examen estudiantes de distintas partes de un país y con diferente CI. Analizar esta tabla al nivel de significancia 0.05 y dar las conclusiones. Usar MINITAB para elaborar la tabla de ANOVA.

16.49 Usar SAS para resolver el problema 16.48 al nivel de significancia 0.01.

16.50 En el problema 16.37, ¿se puede determinar si hay diferencia significativa en la cantidad de artículos producidos en los distintos días de la semana? Explicar.

Tabla 16.49

	Puntuación en el examen		
	CI alto	CI medio	CI bajo
Este	88	80	72
Oeste	84	78	75
Sur	86	82	70
Norte y centro	80	75	79

16.51 Se sabe que en los cálculos del análisis de varianza a cada entrada se le puede sumar o restar una constante adecuada sin que esto afecte las conclusiones. ¿Pasa lo mismo si cada entrada se multiplica o se divide entre una constante adecuada? Justificar la respuesta.

16.52 Deducir las ecuaciones (*24*), (*25*) y (*26*) para cantidades desiguales de observaciones.

16.53 Suponer que los resultados en la tabla 16.46 del problema 16.43 corresponden al noreste de Estados Unidos y que para el oeste, los resultados correspondientes son los dados en la tabla 16.50. Al nivel de significancia 0.05, determinar si hay diferencias que se deban a: *a*) la sustancias químicas y *b*) la ubicación. Usar MINITAB para elaborar la tabla de ANOVA.

Tabla 16.50

A	5	4	6	3
B	3	4	2	3
C	5	7	4	6

Tabla 16.51

A	17	14	18	12
B	20	10	20	15
C	18	15	16	17
D	12	11	14	11
E	15	12	19	14

15.54 Volviendo a los problemas 16.21 y 16.35, suponer que se lleva a cabo otro experimento en el noreste y se obtienen los resultados dados en la tabla 16.51. Al nivel de significancia 0.05, determinar si hay diferencias entre los rendimientos debidas a: *a*) los fertilizantes y *b*) las tres ubicaciones. Emplear STATISTIX para elaborar la tabla de ANOVA.

16.55 Resolver el problema 16.54 al nivel de significancia 0.01. Usar MINITAB para elaborar la tabla de ANOVA.

16.56 Al nivel de significancia 0.05, hacer un análisis de varianza del cuadrado latino de la tabla 16.52 y proporcionar las conclusiones. Usar SPSS para construir la tabla de ANOVA.

Tabla 16.52

	Factor 1		
Factor 2	B 16	C 21	A 15
	A 18	B 23	C 14
	C 15	A 18	B 12

16.57 Estructurar un experimento que lleve al cuadrado latino de la tabla 16.52.

16.58 Al nivel de significancia 0.05, realizar el análisis de varianza del cuadrado grecolatino de la tabla 16.53 y proporcionar las conclusiones. Usar SPSS para elaborar la tabla de ANOVA.

Tabla 16.53

	Factor 1			
Factor 2	A_γ 6	B_β 12	C_δ 4	D_α 18
	B_δ 3	A_α 8	D_γ 15	C_β 14
	D_β 15	C_γ 20	B_α 9	A_δ 5
	C_α 16	D_δ 6	A_β 17	B_γ 7

16.59 Diseñar un experimento que conduzca al cuadrado grecolatino de la tabla 16.53.

16.60 Describir cómo usar el análisis de varianza en un experimento de tres factores con replicaciones.

16.61 Diseñar y resolver un problema que ilustre el procedimiento del problema 16.60.

16.62 Probar: *a*) la ecuación (*30*) y *b*) las ecuaciones (*31*) a (*34*) de este capítulo.

16.63 En la práctica, ¿se esperaría hallar: *a*) un cuadrado latino 2×2 y *b*) un cuadrado grecolatino de 3×3? Explicar.

PRUEBAS NO PARAMÉTRICAS

17

INTRODUCCIÓN

La mayor parte de las pruebas de hipótesis y significancia (o reglas de decisión), vistas en los capítulos anteriores, requieren varias suposiciones acerca de la población de la que se toma la muestra. Por ejemplo, en la clasificación en un sentido del capítulo 16 se requiere que las poblaciones tengan una distribución normal y desviaciones estándar iguales.

En la práctica, hay situaciones en las que tales suposiciones no se justifican o en las que se duda que se satisfagan, como es el caso de poblaciones muy sesgadas. Debido a esto, se han desarrollado diversas pruebas y métodos que son independientes tanto de la distribución de las poblaciones como de sus correspondientes parámetros. Estas pruebas se conocen como *pruebas no paramétricas*.

Las pruebas no paramétricas se emplean como sustitutos sencillos de pruebas más complicadas; son especialmente útiles cuando se tienen datos no numéricos, como en el caso de consumidores que ordenan cereales u otros productos, de acuerdo con su preferencia.

LA PRUEBA DE LOS SIGNOS

Considérese la tabla 17.1 en la que se muestran las cantidades de tornillos defectuosos producidos en 12 días consecutivos con dos máquinas (I y II); se supone que las dos máquinas tienen la misma producción total diaria. Se desea probar la hipótesis H_0 de que no hay diferencia entre las máquinas: que las diferencias observadas entre las máquinas, en términos de cantidades de tornillos defectuosos producidos, son resultado de la casualidad, lo que equivale a decir que las muestras provienen de la misma población.

Tabla 17.1

Día	1	2	3	4	5	6	7	8	9	10	11	12
Máquina I	47	56	54	49	36	48	51	38	61	49	56	52
Máquina II	71	63	45	64	50	55	42	46	53	57	75	60

Una sencilla prueba no paramétrica para muestras por pares es la *prueba de los signos*. Esta prueba consiste en calcular las diferencias entre las cantidades de tornillos defectuosos producidos por día y anotar únicamente el *signo* de cada diferencia, por ejemplo, el día 1 la diferencia es 47-71, que es negativa. De esta manera, a partir de la tabla 17.1 se obtiene la secuencia de signos siguiente.

$$- \quad - \quad + \quad - \quad - \quad - \quad + \quad - \quad + \quad - \quad - \quad - \qquad (1)$$

(es decir, 3 signos más y 9 signos menos). Si es igualmente probable obtener un $+$ que un $-$, se esperaría que se obtuvieran 6 de cada uno. La prueba H_0 es, entonces, equivalente a preguntarse si una moneda está o no cargada, si en 12 lanzamientos de la moneda se obtienen 3 caras ($+$) y 9 cruces ($-$). Esto implica la distribución binomial vista en el capítulo 7. En el problema 17.1 se muestra que empleando una prueba de dos colas con esta distribución, al nivel de significancia 0.05, no se puede rechazar H_0; es decir, a este nivel no hay diferencia entre las máquinas.

Nota 1: Si algún día las máquinas producen la misma cantidad de tornillos defectuosos, una diferencia en la secuencia (*1*) será *cero*. En este caso, se eliminan esos valores muestrales y se usan 11 en vez de 12 observaciones.

Nota 2: También puede usarse una aproximación normal a la distribución binomial empleando la corrección por continuidad (ver problema 17.2).

Aunque la prueba de los signos es especialmente útil para muestras por pares, como la muestra de la tabla 17.1, también puede usarse para problemas con muestras sencillas (no pares) (ver los problemas 17.3 y 17.4).

LA PRUEBA *U* DE MANN-WHITNEY

Considérese la tabla 17.2, en la que se dan las resistencias de cables hechos de dos aleaciones distintas, I y II. En esta tabla se tienen dos muestras: 8 cables de la aleación I, y 10 cables de la aleación II. Se quiere decidir si hay diferencia entre las muestras o, lo que es lo mismo, si provienen o no de la misma población. Aunque este problema se puede resolver empleando la prueba *t* del capítulo 11, también puede utilizarse una prueba no paramétrica llamada la *prueba U de Mann-Whitney*. Esta prueba consta de los pasos siguientes:

Tabla 17.2

Aleación I				Aleación II				
18.3	16.4	22.7	17.8	12.6	14.1	20.5	10.7	15.9
18.9	25.3	16.1	24.2	19.6	12.9	15.2	11.8	14.7

Paso 1. Se combinan todos los valores muestrales, se ordenan de menor a mayor, y a cada uno de los valores se le asigna una posición o rango (en este caso del 1 al 18). Si dos o más valores muestrales son idénticos (es decir, si hay *puntuaciones empatadas*, o *empates*), a cada uno de los valores muestrales se les asigna una posición (o rango) igual a la *media* de las posiciones que les tocaría ocupar. Si en la tabla 17.2 la entrada 18.9 fuera 18.3, las posiciones 12 y 13 estarían ocupadas por dos valores idénticos, de manera que la posición (o rango) asignada a cada uno sería $\frac{1}{2}(12 + 13) = 12.5$.

Paso 2. Se obtiene la suma de los rangos de cada muestra. Estas sumas se denotan R_1 y R_2, siendo N_1 y N_2 los respectivos tamaños muestrales. Por conveniencia se elige como N_1 la muestra más pequeña, si éstas no son iguales, de manera que $N_1 \leq N_2$. Una diferencia significativa entre las sumas de los rangos R_1 y R_2 implica una diferencia significativa entre las muestras.

Paso 3. Para probar la diferencia entre las sumas de los rangos, se usa el estadístico

$$U = N_1 N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 \tag{2}$$

que corresponde a la muestra 1. La distribución muestral de U es simétrica y tiene media y varianza dadas, respectivamente, por las fórmulas

$$\mu_U = \frac{N_1 N_2}{2} \qquad \sigma_U^2 = \frac{N_1 N_2 (N_1 + N_2 + 1)}{12} \tag{3}$$

Si tanto N_1 como N_2 son por lo menos igual a 8, entonces la distribución de U es aproximadamente normal, de manera que

$$z = \frac{U - \mu_U}{\sigma_U} \tag{4}$$

tiene una distribución normal con media 0 y desviación estándar 1. Empleando el apéndice II se puede, entonces, decidir si las muestras son o no significativamente diferentes. En el problema 17.5 se muestra que, al nivel de significancia 0.05, hay una diferencia significativa entre los cables.

Nota 3: Un valor de U correspondiente a la muestra 2 es el dado por el estadístico

$$U = N_1 N_2 + \frac{N_2(N_2 + 1)}{2} - R_2 \tag{5}$$

y tiene la misma distribución muestral que el estadístico (2), siendo su media y su varianza las dadas por la fórmula (3). El estadístico (5) se relaciona con el estadístico (2), ya que si U_1 y U_2 son los valores correspondientes a los estadísticos (2) y (5), respectivamente, entonces se tiene

$$U_1 + U_2 = N_1 N_2 \tag{6}$$

También se tiene

$$R_1 + R_2 = \frac{N(N + 1)}{2} \tag{7}$$

donde $N = N_1 + N_2$. La fórmula (7) puede servir como verificación de los cálculos.

Nota 4: En la ecuación (2), el estadístico U es el número de veces que los valores de la muestra 1 preceden a los valores de la muestra 2, cuando todos los valores han sido ordenados en forma creciente de magnitud. Esto proporciona un *método alternativo de conteo* para hallar U.

LA PRUEBA H DE KRUSKAL-WALLIS

La prueba U es una prueba no paramétrica para decidir si dos muestras provienen o no de una misma población. Una generalización de esta prueba para k muestras es la *prueba H de Kruskal-Wallis* o simplemente *prueba H*.

Esta prueba se puede describir de la manera siguiente: supóngase que se tienen k muestras, cuyos tamaños son $N_1, N_2, \ldots, N_k$, por lo que el tamaño de todas estas muestras juntas será $N = N_1 + N_2 + \cdots + N_k$. Supóngase además que todas estas muestras se juntan y sus valores se ordenan de menor a mayor asignándoles un rango, y que las sumas de los rangos, de cada una de las k muestras, son $R_1, R_2, \ldots, R_k$, respectivamente. Se define el estadístico como

$$H = \frac{12}{N(N + 1)} \sum_{j=1}^{k} \frac{R_j^2}{N_j} - 3(N + 1) \tag{8}$$

se puede demostrar que la distribución muestral de H es aproximadamente una *distribución chi cuadrada* con $k - 1$ grados de libertad, siempre que cada uno de los $N_1, N_2, \ldots, N_k$, sean por lo menos de 5.

La prueba H proporciona un método no paramétrico de *análisis de varianza* para clasificaciones en un sentido o experimentos de un factor, pudiéndose hacer generalizaciones.

PRUEBA H CORREGIDA PARA EMPATES

Cuando hay demasiados empates entre las observaciones de los datos muestrales, el valor de H dado por el estadístico (8) es menor de lo que debiera ser. El valor correcto de H, que se denota H_c, se obtiene dividiendo el valor dado por el estadístico (8) entre el factor de corrección

$$1 - \frac{\sum (T^3 - T)}{N^3 - N} \tag{9}$$

donde T es la cantidad de empates que corresponden a cada observación y donde la suma se toma sobre todas las observaciones. Si no hay empates, entonces $T = 0$ y el factor (*9*) se reduce a 1, de manera que no se necesita ninguna corrección. En la práctica, la corrección suele ser despreciable (es decir, no es suficiente para garantizar que haya un cambio de decisión).

PRUEBA DE LAS RACHAS PARA ALEATORIEDAD

Aunque la palabra "aleatorio" se ha usado muchas veces en este libro, en ninguno de los capítulos anteriores se ha dado una prueba para aleatoriedad. La *teoría de las rachas* proporciona una prueba no paramétrica para aleatoriedad.

Para entender qué es una racha, considérese una secuencia formada por dos símbolos, *a* y *b*, por ejemplo,

$$a\ a\ |\ b\ b\ b\ |\ a\ |\ b\ b\ |\ a\ a\ a\ a\ a\ |\ b\ b\ b\ |\ a\ a\ a\ a\ | \tag{10}$$

Por ejemplo, en el lanzamiento de una moneda, *a* puede representar "cara" y *b* puede ser "cruz"; o al muestrear los tornillos producidos con una máquina, *a* puede corresponder a "defectuoso" y *b* a "no defectuoso".

Una *racha* se define como un conjunto de símbolos idénticos (o semejantes) que se encuentran entre dos símbolos diferentes o entre ningún símbolo (como el principio y el final de la secuencia). Si se lee la secuencia (*10*) de izquierda a derecha, la primera racha, cuyo fin está señalado por una línea vertical, consta de dos *a*; de manera similar, la segunda racha consta de tres *b*; la tercera racha consta de una *a*, etc. En total hay siete rachas.

Parece claro que debe existir alguna relación entre aleatoriedad y cantidad de rachas. Así, en la secuencia

$$a\ |\ b\ |\ a\ |\ b\ |\ a\ |\ b\ |\ a\ |\ b\ |\ a\ |\ b\ |\ a\ |\ b\ | \tag{11}$$

se observa un *patrón cíclico*, en el que aparece una *a* y luego una *b*, otra vez una *a* y luego una *b*, etc., que sería difícil pensar que fuera aleatorio. En este caso, se tienen *demasiadas* rachas (en realidad, se tiene la cantidad máxima posible, dada la cantidad de letras *a* y letras *b*).

Por otro lado, en la secuencia

$$a\ a\ a\ a\ a\ a\ |\ b\ b\ b\ b\ |\ a\ a\ a\ a\ a\ |\ b\ b\ b\ | \tag{12}$$

parece haber un *patrón de tendencia* en el que se agrupan (o acumulan) las letras *a* y las letras *b*. En este caso hay *muy pocas* rachas y no se puede considerar que esta secuencia sea aleatoria.

Por lo tanto, se considerará que una secuencia no es aleatoria si hay demasiadas o muy pocas rachas; si no es así, se considera que la secuencia es aleatoria. Para cuantificar esta idea, supóngase que se forman todas las secuencias posibles que tengan una cantidad N_1 de letras *a* y una cantidad N_2 de letras *b* haciendo un total N de símbolos (o letras) ($N_1 + N_2 = N$). La colección de todas estas secuencias proporciona una distribución muestral: a cada secuencia le corresponde una cantidad de rachas, denotada V. De esta manera se llega a la distribución muestral del estadístico V. Puede demostrarse que esta distribución muestral tiene media y varianza dadas, respectivamente, por las fórmulas

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1 \qquad \sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)} \tag{13}$$

Empleando las fórmulas (*13*), se puede probar la hipótesis de aleatoriedad al nivel de significancia adecuado. Se encuentra que, si tanto N_1 como N_2 son por lo menos 8, entonces la distribución muestral de V se aproxima a una distribución normal. Por lo tanto,

$$z = \frac{V - \mu_V}{\sigma_V} \tag{14}$$

tiene distribución normal, con media 0 y varianza 1, y por lo tanto puede emplearse el apéndice II.

OTRAS APLICACIONES DE LA PRUEBA DE LAS RACHAS

Las siguientes son otras aplicaciones de las rachas para problemas estadísticos:

1. **Prueba mayor y menor que la mediana para aleatoriedad de datos numéricos.** Para determinar si un conjunto de datos numéricos es aleatorio (por ejemplo, los datos de una muestra), primero se colocan los datos en el *mismo orden* en que se obtuvieron. Después, se encuentra la mediana de esos datos y cada dato se reemplaza por una a, si su valor es mayor que la mediana o por una b si su valor es menor que la mediana. Si un valor es igual a la mediana, se elimina de la muestra. La muestra es o no aleatoria según si la secuencia de letras a y b sea o no aleatoria. (Ver problema 17.20.)

2. **Diferencias entre poblaciones de las que se ha tomado una muestra.** Supóngase que dos muestras de tamaños m y n se denotan $a_1, a_2, \ldots, a_m$ y $b_1, b_2, \ldots, b_n$, respectivamente. Para decidir si las muestras provienen o no de una misma población, primero se ordenan todos los $m + n$ valores muestrales de menor a mayor. Si hay valores iguales, se deben ordenar mediante un proceso aleatorio (por ejemplo, empleando números aleatorios). Si la secuencia obtenida es aleatoria, se concluye que las muestras realmente no son diferentes y que, por lo tanto, provienen de la misma población; si la secuencia no es aleatoria, no puede sacarse tal conclusión. Esta prueba puede ser una alternativa a la prueba U de Mann-Whitney. (Ver problema 17.21.)

CORRELACIÓN DE RANGOS DE SPEARMAN

Los métodos no paramétricos también pueden usarse para medir la correlación entre dos variables, X y Y. En lugar de emplear valores precisos de las variables, o cuando no se puede tener tal precisión, los datos se ordenan desde 1 hasta N de acuerdo con su tamaño, importancia, etc. Una vez que las variables X y Y se han ordenado de esta manera, el *coeficiente de correlación de rangos* o *fórmula de Spearman para correlación de rangos* (como suele llamársele) es

$$r_S = 1 - \frac{6 \sum D^2}{N(N^2 - 1)} \tag{15}$$

donde D denota las diferencias entre los rangos de los valores correspondientes de X y Y y donde N es la cantidad de pares de valores (X, Y) que hay en los datos.

PROBLEMAS RESUELTOS

LA PRUEBA DE LOS SIGNOS

17.1 Hágase referencia a la tabla 17.1 y al nivel de significancia 0.05, se prueba la hipótesis H_0 de que no hay diferencia entre las máquinas, contra la hipótesis alternativa H_1 de que sí hay diferencia.

SOLUCIÓN

En la figura 17-1 se muestra la distribución binomial de X caras en 12 lanzamientos de una moneda, en forma de áreas bajo los rectángulos, para $X = 0, 1, \ldots, 12$. Sobrepuesta a la distribución binomial se encuentra la distribución normal, trazada con una línea punteada. La media de la distribución binomial es $\mu = Np = 12(0.5) = 6$. La desviación estándar es $\sigma = \sqrt{Npq} = \sqrt{12(0.5)(0.5)} = \sqrt{3} = 1.73$. La distribución normal también tiene media $= 6$ y desviación estándar $= 1.73$. De acuerdo con el capítulo 7, la probabilidad binomial de X caras es

$$\Pr\{X\} = \binom{12}{X}\left(\frac{1}{2}\right)^X\left(\frac{1}{2}\right)^{12-X} = \binom{12}{X}\left(\frac{1}{2}\right)^{12}$$

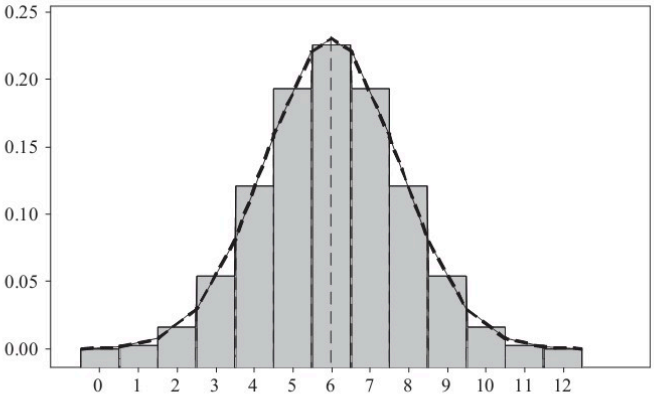


Figura 17-1 Distribución binomial (áreas bajo los rectángulos) y aproximación normal a la distribución binomial (curva punteada).

Las probabilidades pueden encontrarse usando EXCEL. El valor p correspondiente al resultado $X = 3$ es $2P\{X \leq 3\}$, que usando EXCEL es `=2*BINOMDIST(3,12,0.5,1)` o bien 0.146. (El valor p es el doble del área en la cola izquierda de la distribución binomial.) Como esta área es mayor que 0.05, no se puede rechazar la hipótesis nula al nivel de significancia 0.05. Por lo tanto, se concluye que a este nivel no hay diferencia entre las dos máquinas.

17.2 Resolver el problema 17.1, pero esta vez empleando la aproximación normal a la distribución binomial.

SOLUCIÓN

En la aproximación normal a la distribución binomial se emplea el hecho de que la puntuación z correspondiente a la cantidad de caras es

$$z = \frac{X - \mu}{\sigma} = \frac{X - Np}{\sqrt{Npq}}$$

Como en la distribución normal la variable X es discreta, en tanto que en la distribución binomial es continua, se hace una *corrección por continuidad* (por ejemplo, 3 caras es en realidad un valor que está entre 2.5 y 3.5 caras). Esto es equivalente a restarle 0.05 al valor de X si $X > Np$ y sumarle 0.05 al valor de X si $X < Np$. Como $N = 12$, $\mu = Np = (12)(0.5) = 6$ y $\sigma = \sqrt{Npq} = \sqrt{(12)(0.5)(0.5)} = 1.73$, se tiene

$$z = \frac{(3 + 0.5) - 6}{1.73} = -1.45$$

El valor p es el doble del área a la izquierda de -1.45. Empleando EXCEL con `=2*NORMSDIST(-1.45)` se obtiene 1.47. En la figura 17-1, el valor p es, aproximadamente, el área bajo la curva normal estándar a la izquierda de -1.45, la cual se duplica debido a que se trata de una hipótesis de dos colas. Obsérvese cuán cercanos, uno de otro, están los dos valores p; el área binomial bajo los rectángulos es 0.146 y el área bajo la curva normal estándar es 0.147.

17.3 La empresa PQR asegura que un tipo de batería fabricada por ellos tiene una duración mayor a 250 horas (h). Para determinar si está justificado, se mide la duración de 24 baterías producidas por esta empresa; los resultados se presentan en la tabla 17.3. Suponiendo que la muestra sea aleatoria, determinar al nivel de significancia 0.05 si lo que asegura la empresa está justificado. Resolver el problema primero a mano, dando todos los detalles de la prueba de los signos. Después, dar la solución empleando MINITAB.

SOLUCIÓN

Sea H_0 la hipótesis de que la duración de las baterías de esta empresa es igual a 250 h y sea H_1 la hipótesis de que su duración es mayor a 250 h. Para probar H_0 contra H_1 se emplea la prueba de los signos. Para esto, a cada entrada de la tabla 17.3 se le resta 250 y se registra el signo de la diferencia, como se muestra en la tabla 17.4. Se observa que hay 15 signos más y 9 signos menos.

Tabla 17.3

271	230	198	275	282	225	284	219
253	216	262	288	236	291	253	224
264	295	211	252	294	243	272	268

Tabla 17.4

+	−	−	+	+	−	+	−
+	−	+	+	−	+	+	−
+	+	−	+	+	−	+	+

Empleando la aproximación normal a la distribución binomial, la puntuación z es

$$z = \frac{(15 - 0.5) - 24(0.5)}{\sqrt{24(0.5)(0.5)}} = 1.02.$$

Obsérvese que al restar 0.5 de 15 se hace la corrección por continuidad (15 – 0.5) = 14.5. El valor p es el área de la curva normal estándar, a la derecha de 1.02. (Ver la figura 17-2.)

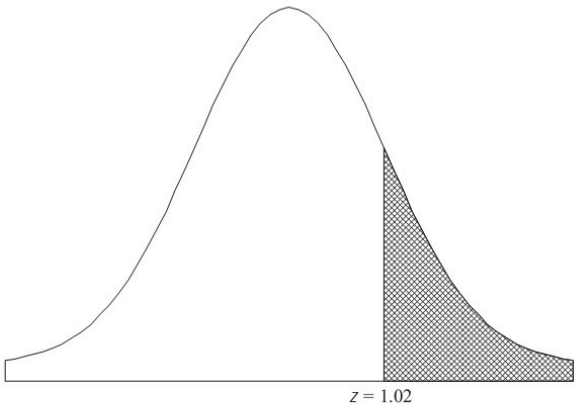


Figura 17-2 El valor p es el área a la derecha de $z = 1.02$.

El valor p es el área a la derecha de $z = 1.02$, o usando EXCEL, el área se obtiene mediante `=1-NORMSDIST(1.02)` o bien 0.1537. Dado que el valor $p > 0.05$, lo que asegura la empresa no se justifica.

A continuación se presenta la solución empleando MINITAB. Los datos de la tabla 17.3 se ingresan en la columna 1 de la hoja de cálculo de MINITAB y a esta columna se le pone como título Duración. Con la secuencia **Stat → Nonparametrics → 1-sample sign** se obtienen los resultados siguientes.

Prueba de los signos para la mediana: Duración

```
Sign test of median = 250.0 versus > 250.0

             N    Below    Equal    Above        P    Median
Lifetime    24        9        0       15   0.1537     257.5
```

Obsérvese que la información dada aquí es la misma a la que se llegó antes en la solución de este problema.

17.4 En la tabla 17.5 se presentan 40 calificaciones obtenidas en un examen a nivel estatal. Al nivel de significancia 0.05, probar la hipótesis de que la calificación mediana de todos los participantes es: *a*) 66 y *b*) 75. Resolver el problema primero a mano, dando todos los detalles de la prueba de los signos, y a continuación, resolverlo empleando MINITAB.

Tabla 17.5

71	67	55	64	82	66	74	58	79	61
78	46	84	93	72	54	78	86	48	52
67	95	70	43	70	73	57	64	60	83
73	40	78	70	64	86	76	62	95	66

SOLUCIÓN

a) Se resta 66 a cada entrada de la tabla 17.5 y se conservan sólo los signos de las diferencias, con lo que se obtiene la tabla 17.6, en la que se observa que hay 23 signos más, 15 signos menos y 2 ceros. Si se eliminan los 2 ceros, la muestra consta de 38 signos: 23 signos más y 15 signos menos. En una prueba de dos colas, usando la distribución normal con probabilidades $\frac{1}{2}(0.05) = 0.025$ en cada cola (ver la figura. 17-3), la regla de decisión que se adopta es la siguiente:

Aceptar la hipótesis si $-1.96 \leq z \leq 1.96$.

Rechazar la hipótesis si no es así.

Dado que
$$z = \frac{X - Np}{\sqrt{Npq}} = \frac{(23 - 0.5) - (38)(0.5)}{\sqrt{(38)(0.5)(0.5)}} = 1.14$$

al nivel de significancia 0.05, se acepta la hipótesis de que la mediana es 66.

Tabla 17.6

+	+	−	−	+	0	+	−	+	−
+	−	+	+	+	−	+	+	−	−
+	+	+	−	+	+	−	−	−	+
+	−	+	+	−	+	+	−	+	0

Obsérvese que también se puede usar 15, la cantidad de signo menos. En este caso

$$z = \frac{(15 + 0.5) - (38)(0.5)}{\sqrt{(38)(0.5)(0.5)}} = -1.14$$

llegándose a la misma conclusión.

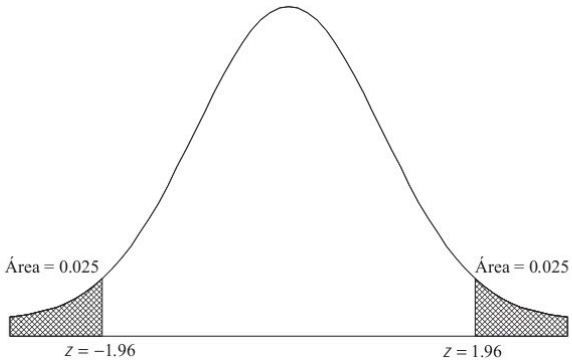


Figura 17-3 Prueba de dos colas mostrando la región crítica para $\alpha = 0.05$.

b) Restando 75 a cada una de las entradas de la tabla 17.5 se obtiene la tabla 17.7, en la que hay 13 signos más y 27 signos menos. Como

$$z = \frac{(13 + 0.5) - (40)(0.5)}{\sqrt{(40)(0.5)(0.5)}} = -2.06$$

al nivel de significancia 0.05, se rechaza la hipótesis de que la mediana sea 75.

Tabla 17.7

−	−	−	−	+	−	−	−	+	−
+	−	+	+	−	−	+	+	−	−
−	+	−	−	−	−	−	−	−	+
−	−	+	−	−	+	+	−	+	−

Empleando este método se puede encontrar un intervalo de confianza de 95% para la calificación mediana en este examen. (Ver el problema 17.30.)

Obsérvese que en la solución anterior se emplea el método clásico de prueba de hipótesis. En el método clásico se emplea $\alpha = 0.05$ para determinar la región de rechazo para la prueba, [$z < -1.96$ o $z > 1.96$]. A continuación se calcula el estadístico de prueba [en el inciso *a*) $z = -1.14$, y en el inciso *b*) $z = -2.06$]. Si el estadístico de prueba cae en la región de rechazo, se rechaza la hipótesis nula. Si este estadístico no cae en la región de rechazo, no se rechaza la hipótesis nula.

En la solución de MINITAB se usa el método del valor *p*. Se calcula el valor *p* y si este valor es menor a 0.05, se rechaza la hipótesis nula. Si el valor *p* es mayor que 0.05, no se rechaza la hipótesis nula. Usando tanto el método clásico como el método del valor *p* se llega a la misma decisión.

La solución empleando MINITAB es la que se muestra a continuación. Para probar que la mediana es 66, el resultado es el siguiente

Prueba de los signos para la mediana: Calificaciones

```
Sign test of median = 66.00 versus not = 66.00

          N   Below   Equal   Above        P   Median
Grade    40      15       2      23   0.2559    70.00
```

Como el valor *p* es mayor que 0.05, no se rechaza la hipótesis nula.
El resultado para probar que la mediana es 75 es

Prueba de los signos para la mediana: Calificaciones

```
Sign test of median = 75.00 versus not = 75.00

          N   Below   Equal   Above        P   Median
Grade    40      27       0      13   0.0385    70.00
```

Como el valor *p* es < 0.05, se rechaza la hipótesis nula.

LA PRUEBA *U* DE MANN-WHITNEY

17.5 Volver a la tabla 17.2. Al nivel de significancia 0.05, determinar si hay alguna diferencia entre los cables hechos con la aleación I y los hechos con la aleación II. Resolver el problema, primero a mano, dando todos los detalles de la prueba *U* de Mann-Whitney, y después usando MINITAB.

SOLUCIÓN

La solución se encuentra siguiendo los pasos 1, 2 y 3 (antes descritos en este capítulo):

Paso 1. Se juntan los 18 valores muestrales y se ordenan de menor a mayor, con lo cual se obtiene el primer renglón de la tabla 17.8. En el segundo renglón se numeran estos valores del 1 al 18, con lo que se obtienen los rangos.

Tabla 17.8

10.7	11.8	12.6	12.9	14.1	14.7	15.2	15.9	16.1	16.4	17.8	18.3	18.9	19.6	20.5	22.7	24.2	25.3
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Paso 2. Para hallar la suma de los rangos de cada muestra, se rescribe la tabla 17.2 con los rangos correspondientes a cada valor de acuerdo con la tabla 17.8, y de esta manera se obtiene la tabla 17.9. La suma de los rangos correspondiente a la aleación I es 106 y la suma de los rangos correspondientes a la aleación II es 65.

Tabla 17.9

Aleación I		Aleación II	
Resistencia del cable	Rango	Resistencia del cable	Rango
18.3	12	12.6	3
16.4	10	14.1	5
22.7	16	20.5	15
17.8	11	10.7	1
18.9	13	15.9	8
25.3	18	19.6	14
16.1	9	12.9	4
24.2	17	15.2	7
	Suma 106	11.8	2
		14.7	6
			Suma 65

Paso 3. Como la muestra de la aleación I es la menor, $N_1 = 8$ y $N_2 = 10$. Las correspondientes sumas de los rangos son $R_1 = 106$ y $R_2 = 65$. Entonces

$$U = N_1N_2 + \frac{N_1(N_1+1)}{2} - R_1 = (8)(10) + \frac{(8)(9)}{2} - 106 = 10$$

$$\mu_U = \frac{N_1N_2}{2} = \frac{(8)(10)}{2} = 40 \qquad \sigma_U^2 = \frac{N_1N_2(N_1+N_2+1)}{12} = \frac{(8)(10)(19)}{12} = 126.67$$

Por lo que $\sigma_U = 11.25$ y

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{10 - 40}{11.25} = -2.67$$

Dado que la hipótesis H_0 que se está probando es que *no* hay diferencia entre las aleaciones, se requiere una prueba de dos colas. La regla de decisión al nivel de significancia 0.05 es:

Aceptar H_0 si $-1.96 \leq z \leq 1.96$.

Rechazar H_0 si no es así.

Como $z = -2.67$, se rechaza H_0 y se concluye que al nivel de significancia 0.05, sí hay diferencia entre las aleaciones.

A continuación se presenta la solución a este problema obtenida con MINITAB. Primero, se ingresan los datos de la aleación I en la columna C1 y los datos de la aleación II en la columna C2, y a las columnas se les pone como encabezado `AleaciónI` y `AleaciónII`. Mediante la secuencia **Stat** → **Nonparametrics** → **Mann-Whitney** se obtienen los resultados siguientes.

Prueba de Mann-Whitney e intervalo de confianza: AleaciónI y AleaciónII

```
           N   Median
AlloyI     8   18.600
AlloyII   10   14.400
```

```
Point estimate for ETA1-ETA2 is     4.800
95.4 Percent CI for ETA1-ETA2 is (2.000, 9.401)
W = 106.0
Test of ETA1 = ETA2  vs  ETA1 not = ETA2 is significant at 0.0088
```

Los resultados de MINITAB dan la resistencia mediana de los cables de cada muestra, una estimación puntual de la diferencia entre las medianas poblacionales (ETA1−ETA2), un intervalo de confianza para la diferencia entre las medianas poblacionales, la suma de los rangos de la primera variable ($W = 106$) y el valor p para dos colas = 0.0088. Como el valor $p < 0.05$, se rechaza la hipótesis nula. Se concluye que con la aleación I se obtienen cables más resistentes.

17.6 Con los datos del problema 17.5, verificar las fórmulas (*6*) y (*7*) de este capítulo.

SOLUCIÓN

a) Como los valores de U que se obtienen con las muestras 1 y 2 son

$$U_1 = N_1N_2 + \frac{N_1(N_1+1)}{2} - R_1 = (8)(10) + \frac{(8)(9)}{2} - 106 = 10$$

$$U_2 = N_1N_2 + \frac{N_2(N_2+1)}{2} - R_2 = (8)(10) + \frac{(10)(11)}{2} - 65 = 70$$

se tiene $U_1 + U_2 = 10 + 70 = 80$ y $N_1N_2 = (8)(10) = 80$.

b) Como $R_1 = 106$ y $R_2 = 65$, se tiene $R_1 + R_2 = 106 + 65 = 171$ y

$$\frac{N(N+1)}{2} = \frac{(N_1+N_2)(N_1+N_2+1)}{2} = \frac{(18)(19)}{2} = 171$$

17.7 Resolver el problema 17.5 usando el estadístico U de la muestra de la aleación II.

SOLUCIÓN

Para la muestra de la aleación II

$$U = N_1N_2 + \frac{N_2(N_2+1)}{2} - R_2 = (8)(10) + \frac{(10)(11)}{2} - 65 = 70$$

de manera que

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{70 - 40}{11.25} = 2.67$$

Este valor de z es el *negativo* de la z del problema 17.5, por lo que se emplea la cola derecha de la distribución normal, en lugar de la cola izquierda. Como este valor de z también se encuentra fuera de $-1.96 \leq z \leq 1.96$, la conclusión es la misma que en el problema 17.5.

17.8 Un profesor tiene dos grupos de psicología: uno en la mañana, con 9 alumnos, y otro en la tarde con 12 alumnos. En el examen final, que es el mismo para los dos grupos, las calificaciones obtenidas son las que se muestran en la tabla 17.10. ¿Puede concluirse al nivel de significancia 0.05 que en el grupo de la mañana el rendimiento sea menor que en el grupo de la tarde? Resolver el problema, primero a mano, dando todos los detalles de la prueba U de Mann-Whitney y después dar la solución empleando MINITAB.

Tabla 17.10

Grupo matutino	73	87	79	75	82	66	95	75	70			
Grupo vespertino	86	81	84	88	90	85	84	92	83	91	53	84

SOLUCIÓN

Paso 1. En la tabla 17.11 se muestran las calificaciones con sus rangos respectivos. Obsérvese que el rango que corresponde a las dos calificaciones 75 es $\frac{1}{2}(5+6) = 5.5$ y el que corresponde a los tres 84 es $\frac{1}{3}(11+12+13) = 12$.

Tabla 17.11

53	66	70	73	75	75	79	81	82	83	84	84	84	85	86	87	88	90	91	92	92
1	2	3	4	5.5		7	8	9	10	12			14	15	16	17	18	19	20	21

Paso 2. Rescribiendo la tabla 17.10 en términos de los rangos, se obtiene la tabla 17.12.
Verificación: $R_1 = 73$, $R_2 = 158$ y $N = N_1 + N_2 = 9 + 12 = 21$; por lo tanto, $R_1 + R_2 = 73 + 158 = 231$ y

$$\frac{N(N+1)}{2} = \frac{(21)(22)}{2} = 231 = R_1 + R_2$$

Tabla 17.12

													Suma de rangos
Grupo matutino	4	16	7	5.5	9	2	21	5.5	3				73
Grupo vespertino	15	8	12	17	18	14	12	20	10	19	1	12	158

Paso 3.

$$U = N_1N_2 + \frac{N_1(N_1+1)}{2} - R_1 = (9)(12) + \frac{(9)(10)}{2} - 73 = 80$$

$$\mu_U = \frac{N_1N_2}{2} = \frac{(9)(12)}{2} = 54 \qquad \sigma_U^2 = \frac{N_1N_2(N_1+N_2+1)}{12} = \frac{(9)(12)(22)}{12} = 198$$

Por lo tanto,

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{80 - 54}{14.07} = 1.85$$

El valor p para una cola se obtiene empleando la expresión de EXCEL `=1-NORMSDIST(1.85)` que da 0.0322. Como el valor $p < 0.05$, se concluye que el desempeño del grupo matutino no es tan bueno como el del turno vespertino.

La solución de MINITAB al problema es como sigue. Primero se introducen los datos de los valores del turno matutino y del vespertino en las columnas `C1` y `C2` y se nombra a esas columnas `matutino` y `vespertino`. Con la secuencia **Stat** → **Nonparametrics** → **Mann-Whitney** se obtienen los resultados siguientes.

Prueba de Mann–Whitney e intervalo de confianza: Matutino, Vespertino

```
            N   Median
Morning     9    75.00
Afternoon  12    84.50

Point estimate for ETA1-ETA2 is -9.00
95.7 Percent CI for ETA1-ETA2 is (-15.00, 2.00)
W = 73.0
Test of ETA1 = ETA2 vs ETA1 < ETA2 is significant at 0.0350
The test is significant at 0.0348 (adjusted for ties)
```

En los resultados de MINITAB se da la calificación mediana de cada muestra, una estimación puntual de la diferencia entre las medianas poblacionales, un intervalo de confianza para la diferencia entre las medianas poblacionales, la suma de los rangos de la primera variable (en este caso, el grupo matutino) y el valor p para una cola que es = 0.0350. Como el valor p es menor que 0.05, se rechaza la hipótesis nula. Se concluye que la clase matutina no tiene tan buen desempeño como la clase vespertina.

17.9 Dados los datos de la tabla 17.13, encontrar U usando: *a*) la fórmula (*2*) de este capítulo y *b*) el método de conteo (descrito en la **Nota 4** de este capítulo).

SOLUCIÓN

a) Ordenando todos los datos juntos, de menor a mayor, y asignándoles un rango del 1 al 5, se obtiene la tabla 17.14. Sustituyendo los datos de la tabla 17.13 por sus rangos correspondientes se obtiene la tabla 17.15, en la que se dan las sumas de los rangos, que son $R_1 = 5$ y $R_2 = 10$. Como $N_1 = 2$ y $N_2 = 3$, el valor U para la muestra 1 es

$$U = N_1 N_2 + \frac{N_1(N_1 + 1)}{2} - R_1 = (2)(3) + \frac{(2)(3)}{2} - 5 = 4$$

De igual manera se encuentra el valor U para la muestra 2, que es $U = 2$.

Tabla 17.13

Muestra 1	22	10	
Muestra 2	17	25	14

Tabla 17.14

Datos	10	14	17	22	25
Rango	1	2	3	4	5

Tabla 17.15

				Suma de los rangos
Muestra 1	4	1		5
Muestra 2	3	5	2	10

b) Sustituyendo los valores de la tabla 17.14 con I o II, dependiendo si el valor pertenece a la muestra 1 o a la muestra 2, el primer renglón de la tabla 17.14 se transforma en

Dato	I	II	II	I	II

En esta tabla se ve que

Número de valores de la muestra 1 que preceden al primer valor de la muestra 2 $= 1$
Número de valores de la muestra 1 que preceden al segundo valor de la muestra 2 $= 1$
Número de valores de la muestra 1 que preceden al tercer valor de la muestra 2 $= 2$
Total $= 4$

Por lo tanto, el valor de U que corresponde a la primera muestra es 4.

De igual manera,

Número de valores de la muestra 2 que preceden al primer valor de la muestra 1 $= 0$
Número de valores de la muestra 2 que preceden al segundo valor de la muestra 1 $= 2$
Total $= 2$

Por lo tanto, el valor de U que corresponde a la segunda muestra es 2.

Obsérvese que como $N_1 = 2$ y $N_2 = 3$, estos valores satisfacen $U_1 + U_2 = N_1 N_2$; es decir, $4 + 2 = 6 = (2)(3) = 6$.

17.10 Una población consta de los valores 7, 12 y 15. De esta población se toman dos muestras sin reposición: la muestra 1 consta de un valor y la muestra 2 consta de dos valores. (Estas dos muestras agotan la población.)

a) Encontrar la distribución muestral de U y su gráfica.
b) Encontrar la media y la varianza de la distribución del inciso *a*).
c) Verificar los resultados hallados en el inciso *b*) empleando la fórmula (*3*) de este capítulo.

SOLUCIÓN

a) Se elige el muestreo sin reposición para evitar empates, los cuales se presentarían si, por ejemplo, el valor 12 apareciera en ambas muestras.

Como se observa en la tabla 17.16, hay $3 \cdot 2 = 6$ posibilidades para elegir las muestras. Debe notarse que también se pueden emplear sólo los rangos 1, 2 y 3, en lugar de 7, 12 y 15. El valor U de la tabla 17.16 es el hallado para la muestra 1, pero si se usa la U para la muestra 2, se obtendrá la misma distribución.

Tabla 17.16

Muestra 1	Muestra 2	U
7	12 15	2
7	15 12	2
12	7 15	1
12	15 7	1
15	7 12	0
15	12 7	0

En la figura 17-4 se muestra una gráfica de esta distribución, en la que f es la frecuencia. También puede graficarse la distribución de probabilidad de U; en ese caso, $\Pr\{U=0\} = \Pr\{U=1\} = \Pr\{U=2\} = \frac{1}{3}$. La gráfica que se requiere es igual a la de la figura 17-4, pero con $\frac{1}{3}$ y $\frac{1}{6}$ en lugar de 1 y 2.

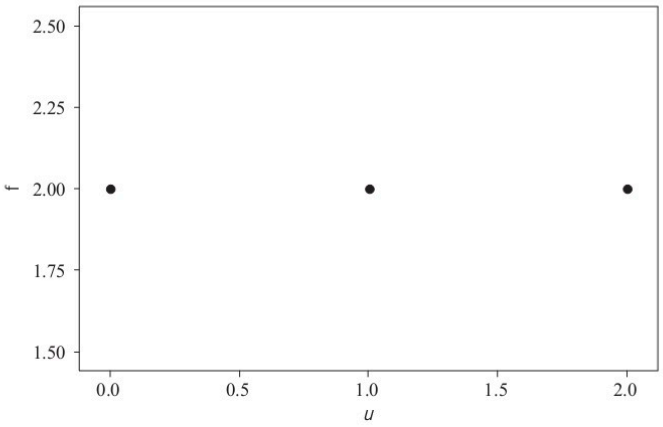


Figura 17-4 MINITAB, gráfica de la distribución muestral de U con $N_1 = 1$ y $N_2 = 2$.

b) La media y la varianza de los valores de la tabla 17.16 son

$$\mu_U = \frac{2+2+1+1+0+0}{6} = 1$$

$$\sigma_U^2 = \frac{(2-1)^2+(2-1)^2+(1-1)^2+(1-1)^2+(0-1)^2+(0-1)^2}{6} = \frac{2}{3}$$

c) De acuerdo con la fórmula (*3*)

$$\mu_U = \frac{N_1 N_2}{2} = \frac{(1)(2)}{2} = 1$$

$$\sigma_U^2 = \frac{N_1 N_2 (N_1 + N_2 + 1)}{12} = \frac{(1)(2)(1+2+1)}{12} = \frac{2}{3}$$

lo que coincide con el inciso *a*).

17.11 *a*) Con los datos del problema 17.9, encontrar la distribución muestral de U y graficarla.
b) Graficar la correspondiente distribución de probabilidad de U.
c) Obtener la media y la varianza de U directamente a partir de los resultados del inciso *a*).
d) Verificar el inciso *c*) empleando la fórmula (*3*) de este capítulo.

SOLUCIÓN

a) En este caso hay $5 \cdot 4 \cdot 3 \cdot 2 = 120$ posibilidades para elegir los valores de las dos muestras y el método del problema 17.9 resulta demasiado laborioso. Para simplificar este procedimiento hay que fijar la atención en la muestra más pequeña (de tamaño $N_1 = 2$) y en las posibles sumas de sus rangos, R_1. La suma de los rangos de la muestra 1 es la *menor* cuando la muestra consta de los dos números de menor rango (1, 2); entonces $R_1 = 1 + 2 = 3$. De igual manera, la suma de los rangos de la muestra 1 es la *mayor* cuando la muestra consta de los dos números de mayor rango (4,5); entonces $R_1 = 4 + 5 = 9$. Por lo tanto, R_1 varía desde 3 hasta 9.

En la columna 1 de la tabla 17.17 se presentan estos valores de R_1 (desde 3 hasta 9) y en la columna 2 se muestran los valores correspondientes de la muestra 1, cuya suma es R_1. En la columna 3 se da la frecuencia (o el número) de las muestras cuya suma es R_1; por ejemplo, hay $f = 2$ muestras para las que $R_1 = 5$. Como $N_1 = 2$ y $N_2 = 3$, se tiene

$$U = N_1 N_2 + \frac{N_1(N_1+1)}{2} - R_1 = (2)(3) + \frac{(2)(3)}{2} - R_1 = 9 - R_1$$

A partir de lo cual pueden encontrarse los valores correspondientes de U en la columna 4 de la tabla; obsérvese que como R_1 varía de 3 a 9, U varía de 6 a 0. La distribución muestral se da en las columnas 3 y 4 y la gráfica en la figura 17-5.

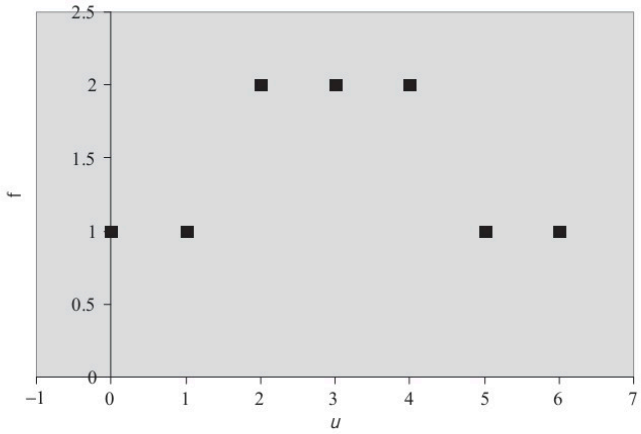


Figura 17-5 EXCEL, gráfica de la distribución muestral de U con $N_1 = 2$ y $N_2 = 3$.

b) La probabilidad de que $U = R_1$ (es decir, $\Pr\{U = R_1\}$) se presenta en la columna 5 de la tabla 17.17 y se obtiene hallando la frecuencia relativa. La frecuencia relativa se halla dividiendo cada frecuencia entre la suma de todas las frecuencias, o sea entre 10; por ejemplo, $\Pr\{U = 5\} = \frac{2}{10} = 0.2$. En la figura 17.6 se muestra la gráfica de la distribución de probabilidad.

Tabla 17.17

R_1	Valores de la muestra 1	f	U	$\Pr\{U = R_1\}$
3	(1, 2)	1	6	0.1
4	(1, 3)	1	5	0.1
5	(1, 4), (2.3)	2	4	0.2
6	(1, 5), (2, 4)	2	3	0.2
7	(2, 5), (3, 4)	2	2	0.2
8	(3, 5)	1	1	0.1
9	(4, 5)	1	0	0.1

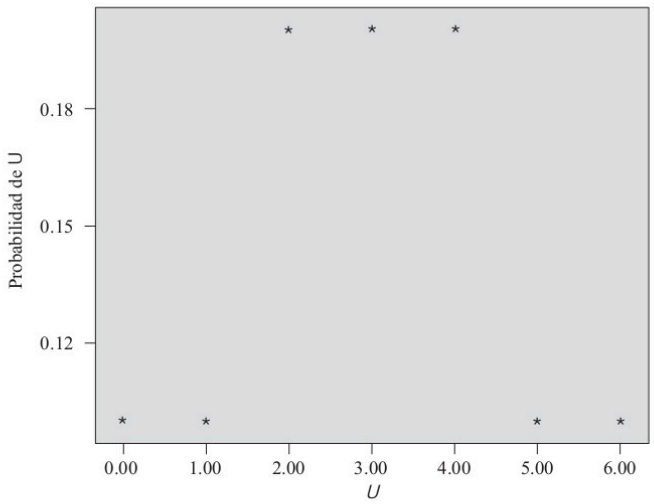


Figura 17-6 SPSS, gráfica de la distribución de probabilidad de U con $N_1 = 2$ y $N_2 = 3$.

c) De acuerdo con las columnas 3 y 4 de la tabla 17.17, se tiene

$$\mu_U = \bar{U} = \frac{\sum fU}{\sum f} = \frac{(1)(6) + (1)(5) + (2)(4) + (2)(3) + (2)(2) + (1)(1) + (1)(0)}{1 + 1 + 2 + 2 + 2 + 1 + 1} = 3$$

$$\sigma_U^2 = \frac{\sum f(U - \bar{U})^2}{\sum f}$$

$$= \frac{(1)(6-3)^2 + (1)(5-3)^2 + (2)(4-3)^2 + (2)(3-3)^2 + (2)(2-3)^2 + (1)(1-3)^2 + (1)(0-3)^2}{10} = 3$$

Otro método

$$\sigma_U^2 = \overline{U^2} - \bar{U}^2 = \frac{(1)(6)^2 + (1)(5)^2 + (2)(4)^2 + (2)(3)^2 + (2)(2)^2 + (1)(1)^2 + (1)(0)^2}{10} - (3)^2 = 3$$

d) De acuerdo con la fórmula (*3*), empleando $N_1 = 2$ y $N_2 = 3$, se tiene

$$\mu_U = \frac{N_1 N_2}{2} = \frac{(2)(3)}{2} = 3 \qquad \sigma_U^2 = \frac{N_1 N_2 (N_1 + N_2 + 1)}{12} = \frac{(2)(3)(6)}{12} = 3$$

17.12 Si N números de un conjunto se ordenan del 1 al N, probar que la suma de los rangos es $[N(N+1)]/2$.

SOLUCIÓN

Sea R la suma de los rangos. Entonces se tiene

$$R = 1 + 2 + 3 + \cdots + (N-1) + N \tag{16}$$

$$R = N + (N-1) + (N-2) + \cdots + 2 + 1 \tag{17}$$

en donde la suma de la ecuación (*17*) se obtiene invirtiendo el orden de los sumandos de la ecuación (*16*). Sumando las ecuaciones (*16*) y (*17*) se obtiene

$$2R = (N+1) + (N+1) + (N+1) + \cdots + (N+1) + (N+1) = N(N+1)$$

como en esta suma $(N+1)$ se presenta N veces, entonces $R = [N(N+1)]/2$. Esto también puede obtenerse empleando una fórmula del álgebra elemental en series y progresiones aritméticas.

17.13 Si R_1 y R_2 son, respectivamente, las sumas de los rangos en las muestras 1 y 2 en una prueba U, demostrar que $R_1 + R_2 = [N(N+1)]/2$.

SOLUCIÓN

Se supone que en los datos muestrales no hay empates. Entonces, R_1 debe ser la suma de algunos de los rangos (números) del conjunto 1, 2, 3, ..., N, y R_2 debe ser la suma de los rangos restantes del conjunto. Por lo tanto, $R_1 + R_2$ debe ser la suma de todos los rangos del conjunto; es decir, $R_1 + R_2 = 1 + 2 + 3 + \cdots + N = [N(N+1)]/2$, de acuerdo con el problema 17.12.

LA PRUEBA *H* DE KRUSKAL-WALLIS

17.14 Una empresa va a comprar una de cinco máquinas: *A*, *B*, *C*, *D* o *E*. En un experimento para determinar si hay diferencia en el desempeño de estas máquinas, cinco operadores experimentados trabajan con cada una de las cinco máquinas durante un mismo tiempo. En la tabla 17.18 se muestra la cantidad de unidades obtenida con cada máquina. A los niveles de significancia: *a*) 0.05 y *b*) 0.01, probar la hipótesis de que no hay diferencia entre las máquinas. Resolver el problema primero a mano, dando todos los detalles de la prueba *H* de Kruskal-Wallis; después, resolver el problema usando MINITAB.

Tabla 17.18

A	68	72	77	42	53
B	72	53	63	53	48
C	60	82	64	75	72
D	48	61	57	64	50
E	64	65	70	68	53

Tabla 17.19

						Suma de los rangos
A	17.5	21	24	1	6.5	70
B	21	6.5	12	6.5	2.5	48.5
C	10	25	14	23	21	93
D	2.5	11	9	14	4	40.5
E	14	16	19	17.5	6.5	73

SOLUCIÓN

Dado que hay cinco muestras (*A*, *B*, *C*, *D* y *E*), $k = 5$. Como cada muestra consta de cinco valores, se tiene $N_1 = N_2 = N_3 = N_4 = N_5 = 5$ y $N = N_1 + N_2 + N_3 + N_4 + N_5 = 25$. Ordenando todos los valores en forma creciente de magnitud y asignando a los empates los rangos adecuados, la tabla 17.18 se transforma en la tabla 17.19, en la que en la columna del extremo derecho se muestran las sumas de los rangos. De acuerdo con la tabla 17.19 se tiene $R_1 = 70$, $R_2 = 48.5$, $R_3 = 93$, $R_4 = 40.5$ y $R_5 = 73$. Por lo tanto,

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{N_j} - 3(N+1)$$

$$= \frac{12}{(25)(26)} \left[\frac{(70)^2}{5} + \frac{(48.5)^2}{5} + \frac{(93)^2}{5} + \frac{(40)^2}{5} + \frac{(73)^2}{5} \right] - 3(26) = 6.44$$

Para $k - 1 = 4$ grados de libertad al nivel de significancia 0.05, en el apéndice IV se encuentra $\chi^2_{.95} = 9.49$. Como $6.44 < 9.49$, al nivel de significancia 0.05, no se puede rechazar la hipótesis de que no haya diferencia entre las máquinas y, por lo tanto, tampoco se podrá rechazar al nivel de significancia 0.01. En otras palabras, se puede aceptar la hipótesis (o postergar la decisión) de que, a ambos niveles, no hay diferencia entre las máquinas.

Obsérvese que este problema ya fue resuelto antes empleando el análisis de varianza (ver problema 16.8) y se llegó a la misma conclusión.

A continuación se presenta la solución del problema usando MINITAB. Primero, es necesario ingresar los datos en la hoja de cálculo de MINITAB. La estructura que deben tener los datos es la siguiente:

Row	Machine	Units
1	1	68
2	1	72
3	1	77
4	1	42
5	1	53
6	2	72
7	2	53
8	2	63
9	2	53
10	2	48
11	3	60
12	3	82
13	3	64
14	3	75
15	3	72
16	4	48
17	4	61
18	4	57
19	4	64
20	4	50
21	5	64
22	5	65
23	5	70
24	5	68
25	5	53

Con la secuencia **Stat** → **Nonparametrics** → **Kruskal-Wallis** se obtienen los resultados:

Prueba de Kruskal-Wallis: unidades *versus* máquinas

Machine	N	Median	Ave Rank	Z
1	5	68.00	14.0	0.34
2	5	53.00	9.7	−1.12
3	5	72.00	18.6	1.90
4	5	57.00	8.1	−1.66
5	5	65.00	14.6	0.54
Overall	25		13.0	

```
H = 6.44   DF = 4   P = 0.168
H = 6.49   DF = 4   P = 0.165 (adjusted for ties)
```

Obsérvese que se dan dos valores *p*. Uno es el valor *p* ajustado, para los empates que se presentan al dar los rangos, y el otro no es ajustado. Es claro que, a los niveles de significancia 0.05 y 0.01, las máquinas no son estadísticamente diferentes respecto a la cantidad de unidades que producen, ya que los dos valores *p* superan con mucho ambos niveles.

17.15 Resolver el problema 17.14 usando los paquetes de software STATISTIX y SPSS.

SOLUCIÓN

El archivo con los datos se estructura como en el problema 17.14. Usando STATISTIX, la secuencia **Statistics → One, Two, Multi-sample tests → Kruskal-Wallis One-way AOV Test** proporciona los resultados que se muestran a continuación. Comparar estos resultados con los obtenidos en el problema 17.14.

```
Statistix 8.0
```

Kruskal-Wallis One-way Nonparametrics AOV para unidades por máquina

Machine	Mean Rank	Sample Size
1	14.0	5
2	9.7	5
3	18.6	5
4	8.1	5
5	14.6	5
Total	13.0	25

Kruskal-Wallis Statistic	6.4949
p-Value, Using Chi-Squared Approximation	0.1651

Los resultados de STATISTIX coinciden con los resultados de MINITAB ajustados para empates.
Los resultados de SPSS también coinciden con los resultados de MINITAB ajustados para empates.

Prueba de Kruskal-Wallis

Rangos

	Máquina	N	Rango medio
Unidades	1.00	4	14.00
	2.00	5	9.70
	3.00	5	18.60
	4.00	5	8.10
	5.00	5	14.60
	Total	25	

Estadísticos de prueba[a,b]

	Unidades
Ji cuadrada	6.495
df	4
Significación asintótica	.165

[a]Prueba de Kruskal-Wallis
[b]Variable agrupada: máquina

17.16 En la tabla 17.20 se da la cantidad de DVD rentados durante el pasado año por los profesores, abogados y médicos de una muestra aleatoria. Usar la prueba *H* de Kruskal-Wallis del paquete SAS para probar la hipótesis nula de que las distribuciones de las rentas son las mismas en los tres grupos de profesionistas. Emplear el nivel 0.01.

Tabla 17.20

Profesores	Abogados	Médicos
18	2	14
4	16	30
5	21	11
9	24	1
20	5	7
26	2	5
7	50	14
17	10	7
43	7	16
20	49	14
24	35	27
7	1	19
34	45	15
30	6	22
45	9	20
2	24	10
45	36	
9	50	
	44	
	3	

SOLUCIÓN

Los datos se ingresan en dos columnas. Una columna contiene 1, 2 o 3 que corresponden a profesores, abogados y médicos; la otra columna contiene la cantidad de DVD rentados. Con la secuencia **Statistics** → **ANOVA** → **Nonparametric Oneway ANOVA** de SAS se obtienen los resultados siguientes.

```
                    The NPAR1WAY Procedure

        Wilcoxon Scores (Rank Sums) for Variable number
              Classified by Variable Profession

                          Sum of      Expected       Std Dev         Mean
Profession       N        Scores     Under H0      Under H0        Score
ffffff
```

El valor p se da como `Pr > Chi-Square 0.6375`. Como el valor p es mucho mayor que 0.05, no se rechaza la hipótesis nula.

PRUEBA DE LAS RACHAS PARA ALEATORIEDAD

17.17 En 30 lanzamientos de una moneda se obtiene la siguiente secuencia de caras (H) y cruces (T):

H T T H T H H H T H H T T H T
H T H H T H T T H T H H T H T

a) Determinar la cantidad V de rachas.

b) Al nivel de significancia 0.05, probar si esta secuencia es aleatoria.

Resolver el problema primero a mano, dando todos los detalles de las pruebas de las rachas para aleatoriedad, y después dar la solución al problema usando MINITAB.

SOLUCIÓN

a) Empleando una línea vertical para separar las rachas

H | T T | H | T | H H H | T | H H | T T | H | T|
H | T | H H | T | H | T T | H | T | H H | T | H | T |

se observa que la cantidad de rachas es $V = 22$.

b) En esta muestra de lanzamientos hay $N_1 = 16$ caras y $N_2 = 14$ cruces, y de acuerdo con el inciso *a*), la cantidad de rachas es $V = 22$. Por lo tanto, de acuerdo con las fórmulas (*13*) de este capítulo, se tiene

$$\mu_V = \frac{2(16)(14)}{16+14} + 1 = 15.93 \qquad \sigma_V^2 = \frac{2(16)(14)[2(16)(14) - 16 - 14]}{(16+14)^2(16+14-1)} = 7.175$$

o $\sigma_V = 2.679$. Por lo tanto, la puntuación z correspondiente a $V = 22$ rachas es

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{22 - 15.93}{2.679} = 2.27$$

En una prueba de dos colas, al nivel de significancia 0.05, se aceptará la hipótesis H_0 de aleatoriedad si $-1.96 \leq z \leq 1.96$ y se rechazará si no es así (ver figura 17-7). Como el valor obtenido para z es $2.27 > 1.96$, se concluye, al nivel de significancia 0.05, que los lanzamientos no son aleatorios. La prueba indica que hay *demasiadas* rachas, lo que indica un *patrón cíclico* en los lanzamientos.

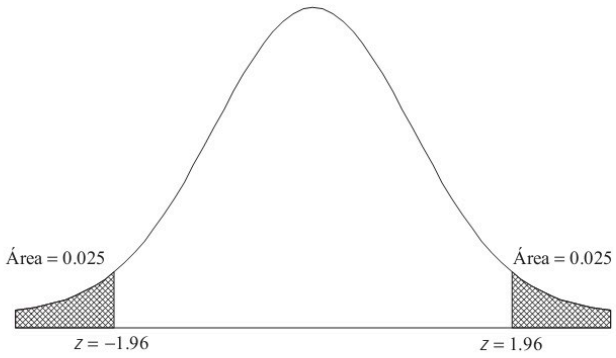


Figura 17-7 Región de rechazo en la curva normal estándar, al nivel de significancia 0.05.

Si se emplea la corrección por continuidad, la puntuación z dada arriba se convierte en

$$z = \frac{(22 - 0.5) - 15.93}{2.679} = 2.08$$

y se llega a la misma conclusión.

La solución del problema empleando MINITAB es como se indica a continuación. En la columna `C1` se ingresan los datos. Cada cara se ingresa como un número 1 y cada cruz como un 0. A la columna 1 se le pone como título `Coin` (moneda). Con la secuencia de MINITAB **Stat** → **Nonparametrics** → **Runs Test** se obtienen los resultados siguientes.

Prueba de las rachas: Moneda

```
Runs test for Coin
Runs above and below K = 0.533333

The observed number of runs = 22
The expected number of runs = 15.9333
16 Observations above K   14 below
p-value = 0.024
```

El valor K corresponde a la media de ceros y unos en la columna 1. La cantidad de observaciones mayores y menores a K es la cantidad de caras y cruces en los 30 lanzamientos de la moneda. El valor p es igual a 0.0235. Como este valor p es menor que 0.05, se rechaza la hipótesis nula. La cantidad de rachas no es aleatoria.

17.18 En una muestra de 48 herramientas producidas con una máquina se encuentra la siguiente secuencia de herramientas buenas (G) y defectuosas (D):

G G G G G G D D G G G G G G G G
G G D D D D G G G G G G D G G G
G G G G G G D D G G G G G D G G

Al nivel de significancia 0.05, probar la aleatoriedad de la secuencia. Usar también SPSS para probar la aleatoriedad de la secuencia.

SOLUCIÓN

El número de letras D es $N_1 = 10$ y el número de letras G es $N_2 = 38$; el número de rachas es $V = 11$. Por lo tanto, la media y la varianza son

$$\mu_V = \frac{2(10)(38)}{10 + 38} + 1 = 16.83 \qquad \sigma_V^2 = \frac{2(10)(38)[2(10)(38) - 10 - 38]}{(10 + 38)^2(10 + 38 - 1)} = 4.997$$

de manera que $\sigma_V = 2.235$.

En una prueba de dos colas, al nivel de significancia 0.05, se acepta la hipótesis H_0 de aleatoriedad si $-1.96 \leq z \leq 1.96$ (ver figura 17-7), y se rechaza si no es así. Como la puntuación z que corresponde a $V = 11$ es

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{11 - 16.83}{2.235} = -2.61$$

y $-2.61 < -1.96$, se rechaza H_0 al nivel 0.05.

La prueba muestra que hay *muy pocas* rachas, lo que indica que hay una acumulación (o agrupación) de herramientas defectuosas. En otras palabras, parece haber un *patrón de tendencia* en la producción de herramientas defectuosas. Se recomienda un examen del proceso de producción.

Con la secuencia **Analyze** → **Nonparametric Tests** → **Runs** de SPSS se obtienen los resultados que se dan a continuación. Si las G se reemplazan por unos y las D por ceros, el valor de prueba (Test Value), 0.7917, es la media de estos valores. Los demás valores que se dan en los resultados son $N_1 = 10$, $N_2 = 38$, Suma = 48 y $V = 11$. El valor obtenido para z es el valor con corrección por continuidad. A esto se debe que el valor z difiera del valor z calculado arriba. El término `Asymp. Sig. (2-tailed)` es el valor p para dos colas correspondiente a $z = -2.386$. Como se ve, los resultados de SPSS contienen la misma información que se halló a mano. Sólo hay que saber interpretarla.

Prueba de rachas

	Calidad
Valor de prueba[a]	.7917
Casos < Valor de prueba	10
Casos > = Valor de prueba	38
Total de casos	48
Número de rachas	11
Z	–2.386
Asymp. Sig. (2 colas)	.017

[a]Media

17.19 *a*) Formar todas las secuencias posibles que contengan tres *a* y dos *b*, y dar la cantidad V de rachas correspondientes a cada secuencia.

b) Obtener la distribución muestral de V y su gráfica.

c) Obtener la distribución de probabilidad de V y su gráfica.

SOLUCIÓN

a) La cantidad de secuencias con tres *a* y dos *b* es

$$\binom{5}{2} = \frac{5!}{2!3!} = 10$$

Estas secuencias se muestran en la tabla 17.21, dando también la cantidad de rachas correspondiente a cada secuencia.

b) En la tabla 17.22 se da la distribución muestral de V (obtenida a partir de la tabla 17.21), en la que V denota la cantidad de rachas y f su frecuencia. Así, por ejemplo, la tabla 17.22 indica que hay un 5, cuatro 4, etc. En la figura 17-8 se muestra la gráfica correspondiente.

Tabla 17.21

Secuencia					Rachas (V)
a	*a*	*a*	*b*	*b*	2
a	*a*	*b*	*a*	*b*	4
a	*a*	*b*	*b*	*a*	3
a	*b*	*a*	*b*	*a*	5
a	*b*	*b*	*a*	*a*	3
a	*b*	*a*	*a*	*b*	4
b	*b*	*a*	*a*	*a*	2
b	*a*	*b*	*a*	*a*	4
b	*a*	*a*	*a*	*b*	3
b	*a*	*a*	*b*	*a*	4

Tabla 17.22

V	f
2	2
3	3
4	4
5	1

c) La distribución de probabilidad de V, graficada en la figura 17-9, se obtiene de la tabla 17.22 dividiendo cada frecuencia entre la frecuencia total $2 + 3 + 4 + 1 = 10$. Por ejemplo, $\Pr\{V = 5\} = \frac{1}{10} = 0.1$.

17.20 En el problema 17.19, obtener directamente de los resultados ahí obtenidos: *a*) la media y *b*) la varianza de la cantidad de rachas.

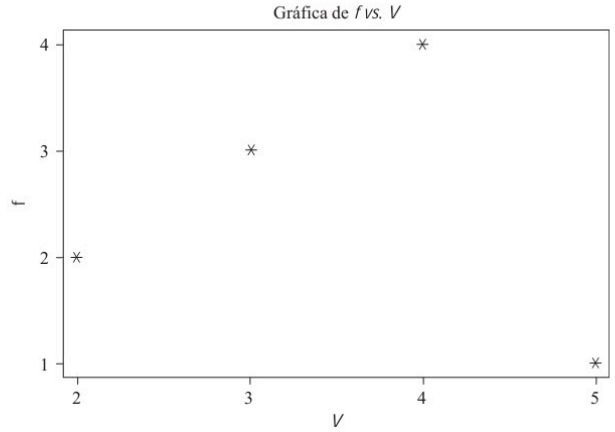


Figura 17-8 STATISTIX, gráfica de la distribución muestral de *V*.

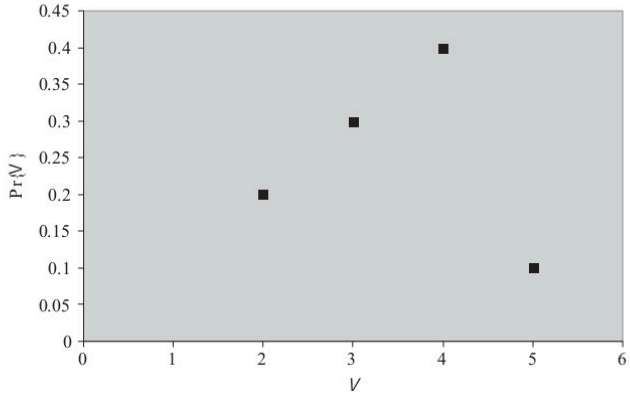


Figura 17-9 EXCEL, gráfica de la distribución de probabilidad de *V*.

SOLUCIÓN

a) De acuerdo con la tabla 17.21, se tiene

$$\mu_V = \frac{2+4+3+5+3+4+2+4+3+4}{10} = \frac{17}{5}$$

Otro método

De acuerdo con la tabla 17.21, con el método de los datos agrupados se tiene

$$\mu_V = \frac{\sum fV}{\sum f} = \frac{(2)(2)+(3)(3)+(4)(4)+(1)(5)}{2+3+4+1} = \frac{17}{5}$$

b) Empleando el método de los datos agrupados para calcular la varianza, de acuerdo con la tabla 17.22, se tiene

$$\sigma_V^2 = \frac{\sum f(V-\bar{V})^2}{\sum f} = \frac{1}{10}\left[(2)\left(2-\frac{17}{5}\right)^2 + (3)\left(3-\frac{17}{5}\right)^2 + (4)\left(4-\frac{17}{5}\right)^2 + (1)\left(5-\frac{17}{5}\right)^2\right] = \frac{21}{25}$$

Otro método

Como en el capítulo 3, la varianza está dada por

$$\sigma_V^2 = \overline{V^2} - \bar{V}^2 = \frac{(2)(2)^2 + (3)(3)^2 + (4)(4)^2 + (1)(5)^2}{10} - \left(\frac{17}{5}\right)^2 = \frac{21}{25}$$

17.21 Repetir el problema 17.20 empleando las fórmulas (*13*) de este capítulo.

SOLUCIÓN

Como hay tres *a* y dos *b*, se tiene que $N_1 = 3$ y $N_2 = 2$. Por lo tanto,

a) $$\mu_V = \frac{2N_1N_2}{N_1+N_2} + 1 = \frac{2(3)(2)}{3+2} + 1 = \frac{17}{5}$$

b) $$\sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1+N_2)^2(N_1+N_2-1)} = \frac{2(3)(2)[2(3)(2)-3-2]}{(3+2)^2(3+2-1)} = \frac{21}{25}$$

OTRAS APLICACIONES DE LA PRUEBA DE LAS RACHAS

17.22 Con los datos del problema 17.3 y empleando como nivel de significancia 0.05, determinar si las duraciones muestrales de las baterías producidas por la empresa PQR son aleatorias. Suponer que las duraciones de las baterías dadas en la tabla 17.3 se registraron en forma consecutiva. Esto es, la primera duración fue 271, la segunda duración fue 230, y así sucesivamente hasta la última duración, 268. Resolver el problema primero a mano, dando todos los detalles de la prueba de las rachas para aleatoriedad. Después, resolver el problema empleando STATISTIX.

SOLUCIÓN

En la tabla 17.23 se muestran las duraciones de las baterías en orden creciente de magnitud. Como en esta tabla hay 24 entradas, la mediana se obtiene de los dos valores de en medio, 253 y 262, y es $\frac{1}{2}(253 + 262) = 257.5$. Ahora se rescriben los datos de la tabla 17.3 sustituyéndolos por una *a* si su valor es mayor a la mediana y por una *b* si su valor es menor a la mediana; se obtiene la tabla 17.24, en la que hay 12 letras *a*, 12 letras *b* y 15 rachas. Por lo tanto, $N_1 = 12$, $N_2 = 12$, $N = 24$, $V = 15$, y se tiene

$$\mu_V = \frac{2N_1N_2}{N_1+N_2} + 1 = \frac{2(12)(12)}{12+12} + 1 = 13 \qquad \sigma_V^2 = \frac{2(12)(12)(264)}{(24)^2(23)} = 5.739$$

de manera que $$z = \frac{V-\mu_V}{\sigma_V} = \frac{15-13}{2.396} = 0.835$$

Tabla 17.23

198	211	216	219	224	225	230	236
243	252	253	253	262	264	268	271
272	275	282	284	288	291	294	295

Tabla 17.24

a	*b*	*b*	*a*	*a*	*b*	*a*	*b*
b	*b*	*a*	*a*	*b*	*a*	*b*	*b*
a	*a*	*b*	*b*	*a*	*b*	*a*	*a*

Empleando una prueba de dos colas, al nivel de significancia 0.05, la hipótesis de aleatoriedad se acepta si $-1.96 \leq z \leq 1.96$. Como 0.835 cae dentro de este intervalo, se concluye que la muestra es aleatoria.

A continuación se presenta el análisis empleando STATISTIX. Las duraciones se ingresan en la columna 1 en el orden en que fueron obtenidas. A la columna se le da como título `Duraciones`. Empleando la secuencia **Statistics** → **Randomness/Normality Tests** → **Runs Test** se obtiene el resultado que se presenta a continuación.

```
Statistix 8.0
```

Prueba de rachas para duración

```
Median                              257.50
Values Above the Median                 12
Values below the Median                 12
Values Tied with the Median              0
Runs Above the Median                    8
Runs Below the Median                    7
Total Number of Runs                    15
Expected Number of Runs               13.0

p-Value, Two-Tailed Test                      0.5264
Probability of getting 15 or fewer runs       0.8496
Probability of getting 14 or more runs        0.2632
```

El valor p grande indica que la cantidad de rachas puede ser considerada como aleatoria.

17.23 Resolver el problema 17.5 empleando la prueba de las rachas para aleatoriedad.

SOLUCIÓN

En el primer renglón de la tabla 17.8 aparecen ya todos los valores de las dos muestras ordenados de menor a mayor. Empleando una a para cada dato de la muestra I y una b para cada dato de la muestra II, el primer renglón de la tabla 17.8 será

$$b \quad b \quad b \quad b \quad b \quad b \quad b \quad b \quad a \quad a \quad a \quad a \quad a \quad b \quad b \quad a \quad a \quad a$$

Como hay cuatro rachas, se tiene $V = 4$, $N_1 = 8$ y $N_2 = 10$. Entonces,

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1 = \frac{2(8)(10)}{18} + 1 = 9.889$$

$$\sigma_V^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1 + N_2)^2(N_1 + N_2 - 1)} + \frac{2(8)(10)(142)}{(18)^2(17)} = 4.125$$

de manera que

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{4 - 9.889}{2.031} = -2.90$$

Si H_0 es la hipótesis de que no hay diferencia entre las aleaciones, esto equivale a la hipótesis de que la secuencia anterior es aleatoria. Esta hipótesis se acepta si $-1.96 \leq z \leq 1.96$ y se rechaza si no es así. Como $z = -2.90$ se encuentra fuera de este intervalo, se rechaza H_0 y se llega a la misma conclusión que en el problema 17.5.

Obsérvese que si se hace la corrección por continuidad,

$$z = \frac{V - \mu_V}{\sigma_V} = \frac{(4 + 0.5) - 9.889}{2.031} = -2.65$$

y se llega a la misma conclusión.

CORRELACIÓN DE RANGOS

17.24 Las calificaciones, en teoría y laboratorio, de una clase de biología que se presentan en la tabla 17.25 corresponden a 10 estudiantes colocados en orden alfabético. Encontrar el coeficiente de correlación de rangos. Usar SPSS para calcular la correlación de rangos de Spearman.

Tabla 17.25

Laboratorio	8	3	9	2	7	10	4	6	1	5
Teoría	9	5	10	1	8	7	3	4	2	6

SOLUCIÓN

En la tabla 17.26 se presentan las diferencias, D, entre los rangos (calificaciones) de laboratorio y de teoría de cada estudiante; se da también D^2 y $\sum D^2$. Entonces

$$r_S = 1 - \frac{6 \sum D^2}{N(N^2 - 1)} = 1 - \frac{6(24)}{10(10^2 - 1)} = 0.8545$$

lo que indica que hay una estrecha relación entre las calificaciones de teoría y de laboratorio.

Tabla 17.26

Diferencia entre los rangos (D)	−1	−2	−1	1	−1	3	1	2	−1	−1	
D^2	1	4	1	1	1	9	1	4	1	1	$\sum D^2 = 24$

Los datos de la tabla 17.26 se ingresan en las columnas tituladas `Lab` y `Lecture` (Teoría). Con la secuencia **Analyze** → **Correlate** → **Bivariate** se obtienen los resultados siguientes.

Correlaciones

			Lab	Teoría
Correlación de rangos de Spearman	Lab	Coeficiente de correlación	1.000	.855
		Sig. (2 colas)	.	.002
		N	10	10
	Teoría	Coeficiente de correlación	.855	1.000
		Sig. (2 colas)	.002	
		N	10	10

El resultado indica que hay una correlación significativa entre los desempeños en `teoría` y en `laboratorio`.

17.25 En la tabla 17.27 se muestran las estaturas de una muestra de 12 padres y de sus hijos mayores adultos. Encontrar el coeficiente de correlación de rangos. Resolver el problema primero a mano, dando todos los detalles para hallar el coeficiente de correlación de rangos. Después, usar SAS para hallar la solución del problema.

Tabla 17.27

Estatura del padre (pulgadas)	65	63	67	64	68	62	70	66	68	67	69	71
Estatura del hijo (pulgadas)	68	66	68	65	69	66	68	65	71	67	68	70

SOLUCIÓN

En orden creciente de magnitud, las estaturas de los padres son

$$62 \quad 63 \quad 64 \quad 65 \quad 66 \quad 67 \quad 67 \quad 68 \quad 68 \quad 69 \quad 71 \tag{18}$$

Como la sexta y séptima estaturas son iguales [67 pulgadas (in)], a estas posiciones se les asigna un rango medio $\frac{1}{2}(6+7) = 6.5$. De igual manera, a las estaturas octava y novena se les asigna el rango $\frac{1}{2}(8+9) = 8.5$. De esta manera, los rangos asignados a las estaturas de los padres son

$$1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6.5 \quad 6.5 \quad 8.5 \quad 8.5 \quad 10 \quad 11 \quad 12 \tag{19}$$

En forma análoga, en orden creciente de magnitud, las estaturas de los hijos son

$$65 \quad 65 \quad 66 \quad 66 \quad 67 \quad 68 \quad 68 \quad 68 \quad 68 \quad 69 \quad 70 \quad 71 \tag{20}$$

y como las estaturas sexta, séptima, octava y novena son iguales (68 in), a estos lugares se les asigna el rango medio $\frac{1}{4}(6+7+8+9) = 7.5$. De esta manera, los rangos asignados a las estaturas de los hijos son

$$1.5 \quad 1.5 \quad 3.5 \quad 3.5 \quad 5 \quad 7.5 \quad 7.5 \quad 7.5 \quad 7.5 \quad 10 \quad 11 \quad 12 \tag{21}$$

Empleando las correspondencias (*18*) y (*19*), y (*20*) y (*21*), la tabla 17.27 puede reemplazarse por la tabla 17.28. En la tabla 17.29 se dan las diferencias entre los rangos D, D^2 y $\sum D^2$, con lo que

$$r_S = 1 - \frac{6\sum D^2}{N(N^2-1)} = 1 - \frac{6(72.50)}{12(12^2-1)} = 0.7465$$

Tabla 17.28

Rango del padre	4	2	6.5	3	8.5	1	11	5	8.5	6.5	10	12
Rango del hijo	7.5	3.5	7.5	1.5	10	3.5	7.5	1.5	12	5	7.5	11

Tabla 17.29

D	−3.5	−1.5	−1.0	1.5	−1.5	−2.5	3.5	3.5	−3.5	1.5	2.5	1.0	
D^2	12.25	2.25	1.00	2.25	2.25	6.25	12.25	12.25	12.25	2.25	6.25	1.00	$\sum D^2 = 72.50$

Este resultado coincide con el coeficiente de correlación obtenido mediante otros métodos.

Las estaturas de los padres se ingresan en la primera columna y las de los hijos en la segunda columna y se pulsa el comando de SAS. Con la secuencia **Statistics** → **Descriptive** → **Correlations** se obtienen los resultados siguientes.

```
                    The CORR Procedure
                2 Variables:   Fatherht Sonht

                    Simple Statistics

Variable      N          Mean       Std Dev        Median       Minimum       Maximum
 Fatherht    12      66.66667       2.77434      67.00000      62.00000      71.00000
 Sonht       12      67.58333       1.88092      68.00000      65.00000      71.00000
```

```
Spearman Correlation Coefficient, N = 12
      prob > |r| under H0: Rho = 0

                Fatherht     Sonht
Fatherht        1.00000     0.74026
                            0.0059
Sonht           0.74026     1.00000
                            0.0059
```

Además de los estadísticos sencillos, SAS da el coeficiente de correlación de Spearman, que es 0.74. El valor p, 0.0059, puede emplearse para probar la hipótesis nula de que el coeficiente poblacional de correlación de rangos es igual a 0 *versus* la hipótesis alternativa de que el coeficiente poblacional de correlación de rangos es diferente de 0. Se concluye que en la población sí existe una relación entre estaturas de los padres y estaturas de los hijos.

PROBLEMAS SUPLEMENTARIOS

LA PRUEBA DE LOS SIGNOS

17.26 Una empresa asegura que si sus productos se adicionan al tanque de gasolina de los automóviles, su rendimiento, en millas por galón, aumenta. Para probar esto, se eligen 15 automóviles diferentes y se mide el rendimiento de la gasolina, en millas por galón, con el aditivo y sin éste; los resultados se muestran en la tabla 17.30. Suponiendo que las condiciones de manejo sean las mismas, determinar los niveles de significancia: *a*) 0.05 y *b*) 0.01, si hay alguna diferencia atribuible al uso del aditivo.

Tabla 17.30

Con aditivo	34.7	28.3	19.6	25.1	15.7	24.5	28.7	23.5	27.7	32.1	29.6	22.4	25.7	28.1	24.3
Sin aditivo	31.4	27.2	20.4	24.6	14.9	22.3	26.8	24.1	26.2	31.4	28.8	23.1	24.0	27.3	22.9

17.27 ¿Se puede concluir, al nivel de significancia 0.05, que el rendimiento, en millas por galón, obtenido en el problema 17.26 es *mejor* con aditivo que sin aditivo?

17.28 Un club para bajar de peso anuncia un programa especial con el que se pierde por lo menos el 6% en 1 mes, si se sigue el programa al pie de la letra. Para probar si esto es así, 36 adultos se someten al programa. De éstos, 25 logran bajar la cantidad deseada, 6 aumentan de peso y el resto permanece esencialmente igual. Al nivel de significancia 0.05, determinar si el programa es efectivo.

17.29 Un capacitador asegura que si el personal de ventas de la empresa toma un curso especial, las ventas anuales de la empresa aumentarán. Para probar esto, 24 personas toman el curso. Las ventas de 16 de ellas aumentan, las ventas de 6 de ellas disminuyen y las de 2 de ellas permanecen igual. Al nivel de significancia 0.05, probar la hipótesis de que el curso incrementa las ventas de la empresa.

17.30 La empresa refresquera MW realiza "pruebas de sabor" en 27 lugares en todo Estados Unidos, con objeto de determinar las preferencias del público respecto a dos marcas de cola, *A* y *B*. En ocho lugares se prefirió la marca *A* a la marca *B*, en 17 lugares se prefirió la marca *B* a la marca *A*, y en el resto de los lugares fueron indiferentes. Al nivel de significancia 0.05, ¿se puede concluir que la marca *B* se prefiere a la marca *A*?

17.31 En la tabla 17.31 se muestra la resistencia a la ruptura de una muestra de 25 cuerdas fabricadas por una empresa. Basándose en esta muestra, probar al nivel de significancia 0.05 la afirmación del fabricante de que la resistencia a la ruptura de las cuerdas es: *a*) 25, *b*) 30, *c*) 35 y *d*) 40.

Tabla 17.31

41	28	35	38	23
37	32	24	46	30
25	36	22	41	37
43	27	34	27	36
42	33	28	31	24

17.32 Mostrar cómo obtener límites de confianza de 95% para los datos del problema 17.4.

17.33 Diseñe y resuelva un problema en el que se emplee la prueba de los signos.

LA PRUEBA *U* DE MANN-WHITNEY

17.34 Dos profesores, *A* y *B*, imparten un mismo curso en la universidad XYZ. En la tabla 17.32 se presentan las calificaciones que obtienen sus alumnos en el examen final, que es común para los dos grupos. Al nivel de significancia 0.05, probar la hipótesis de que no hay diferencia entre las calificaciones de los dos profesores.

Tabla 17.32

A	88	75	92	71	63	84	55	64	82	96			
B	72	65	84	53	76	80	51	60	57	85	94	87	73	61

17.35 Volviendo al problema 17.34, al nivel de significancia 0.01, ¿puede concluirse que los alumnos del maestro *B* sean mejores que los alumnos del maestro *A*?

17.36 Un agricultor quiere determinar si hay diferencia en el rendimiento de dos variedades de trigo, I y II. En la tabla 17.33 se muestra la producción de trigo por unidad de área con cada una de las dos variedades de trigo. A los niveles de significancia: *a*) 0.05 y *b*) 0.01, ¿puede concluirse que existe alguna diferencia?

Tabla 17.33

Trigo I	15.9	15.3	16.4	14.9	15.3	16.0	14.6	15.3	14.5	16.6	16.0
Trigo II	16.4	16.8	17.1	16.9	18.0	15.6	18.1	17.2	15.4		

17.37 ¿Puede el agricultor del problema 17.36 concluir que con el trigo II se obtiene mayor rendimiento?

17.38 Una empresa desea determinar si hay diferencia entre dos marcas de gasolina, *A* y *B*. En la tabla 17.34 se muestran las distancias obtenidas por galón con cada una de las dos marcas. Al nivel de significancia 0.05, ¿puede concluirse: *a*) que sí hay diferencia entre las marcas y *b*) que la marca *B* es mejor que la marca *A*?

Tabla 17.34

A	30.4	28.7	29.2	32.5	31.7	29.5	30.8	31.1	30.7	31.8
B	33.5	29.8	30.1	31.4	33.8	30.9	31.3	29.6	32.8	33.0

17.39 ¿Puede emplearse una prueba *U* para determinar si hay diferencia entre las máquinas I y II de la tabla 17.1? Explicar.

17.40 Diseñar y resolver un problema usando la prueba U.

17.41 Dados los datos de la tabla 17.35, encontrar U usando: *a*) el método de la fórmula y *b*) el método del conteo.

17.42 Repetir el problema 17.41 con los datos de la tabla 17.36.

Tabla 17.35

Muestra 1	15	25
Muestra 2	20	32

Tabla 17.36

Muestra 1	40	27	30	56
Muestra 2	10	35		

17.43 Una población consta de los números 2, 5, 9 y 12. De esta población se toman dos muestras, la primera consta de uno de estos números y la segunda consta de los otros tres números.

a) Obtener la distribución muestral de U y su gráfica.
b) Obtener la media y la varianza de esta distribución, tanto directamente como empleando la fórmula.

17.44 Probar que $U_1 + U_2 = N_1 N_2$.

17.45 Probar que $R_1 + R_2 = [N(N+1)]/2$ en el caso en que la cantidad de empates sea: *a*) 1, *b*) 2 y *c*) cualquier número.

17.46 Si $N_1 = 14$, $N_2 = 12$ y $R_1 = 105$, encontrar: *a*) R_2, *b*) U_1 y *c*) U_2.

17.47 Si $N_1 = 10$, $N_2 = 16$ y $U_2 = 60$, encontrar: *a*) R_1, *b*) R_2 y *c*) U_1.

17.48 ¿Cuál es la mayor cantidad de los valores N_1, N_2, R_1, R_2, U_1 y U_2 que se puede determinar a partir de los restantes? Probar la respuesta.

LA PRUEBA *H* DE KRUSKAL-WALLIS

17.49 Se hace un experimento para determinar el rendimiento de cinco variedades de trigo: *A*, *B*, *C*, *D* y *E*. A cada variedad se le asignan cuatro parcelas. En la tabla 17.37 se presentan los rendimientos (en bushels por acre). Suponiendo que en todas las parcelas la fertilidad sea semejante y que las variedades se hayan asignado de manera aleatoria a las parcelas, determinar, a los niveles de significancia: *a*) 0.05 y *b*) 0.01, si hay diferencia significativa entre los rendimientos.

Tabla 17.37

A	20	12	15	19
B	17	14	12	15
C	23	16	18	14
D	15	17	20	12
E	21	14	17	18

Tabla 17.38

A	33	38	36	40	31	35
B	32	40	42	38	30	34
C	31	37	35	33	34	30
D	27	33	32	29	31	28

17.50 Una empresa desea probar cuatro tipos diferentes de neumáticos: *A*, *B*, *C* y *D*. En la tabla 17.38 se dan las duraciones de los neumáticos (en miles de millas) determinadas de acuerdo con su dibujo; cada tipo de neumático se probó en seis automóviles similares asignados aleatoriamente a los neumáticos. A los niveles: *a*) 0.05 y *b*) 0.01, determinar si hay diferencia significativa entre los neumáticos.

17.51 Un maestro desea probar tres métodos de enseñanza: I, II y III. Para esto, elige en forma aleatoria tres grupos de cinco estudiantes cada uno y en cada grupo prueba uno de los métodos de enseñanza. A todos los estudiantes les pone el mismo examen. En la tabla 17.39 se presentan las calificaciones obtenidas. A los niveles de significancia: *a*) 0.05 y *b*) 0.01, determinar si hay diferencia entre estos métodos de enseñanza.

Tabla 17.39

Método I	78	62	71	58	73
Método II	76	85	77	90	87
Método III	74	79	60	75	80

17.52 En la tabla 17.40 se presentan las calificaciones obtenidas por un estudiante durante un semestre en varias materias. A los niveles de significancia: *a*) 0.05 y *b*) 0.01, probar si hay diferencia entre las calificaciones en estas materias.

Tabla 17.40

Matemáticas	72	80	83	75	
Ciencias	81	74	77		
Inglés	88	82	90	87	80
Economía	74	71	77	70	

17.53 Usando la prueba *H*, resolver: *a*) el problema 16.9, *b*) el problema 16.21 y *c*) el problema 16.22.

17.54 Usando la prueba *H*, resolver: *a*) el problema 16.23, *b*) el problema 16.24 y *c*) el problema 16.25.

PRUEBA DE LAS RACHAS PARA ALEATORIEDAD

17.55 En cada una de estas secuencias, determinar la cantidad, *V*, de rachas.

a) A B A B B A A A B B A B

b) H H T H H H T T T T H H T H H T H T

17.56 A 25 personas se les preguntó si les gustaba un producto (lo que se indica por Y y N, respectivamente). El resultado muestral obtenido es el que se presenta en la secuencia siguiente:

Y Y N N N N Y Y Y N Y N N Y N N N N N Y Y Y Y N N

a) Determinar la cantidad, *V*, de rachas.

b) Al nivel de significancia 0.05, probar si estas respuestas son aleatorias.

17.57 Aplicar la prueba de las rachas a las secuencias (*10*) y (*11*) de este capítulo y dar las conclusiones acerca de la aleatoriedad.

17.58 *a*) Formar todas las secuencias posibles que contengan dos letras *a* y una letra *b* y dar el número *V* de rachas correspondiente a cada secuencia.

b) Obtener la distribución muestral de *V*, así como su gráfica.

c) Obtener la distribución de probabilidad de *V*, así como su gráfica.

17.59 En el problema 17.58, encontrar la media y la varianza de *V*: *a*) directamente a partir de la distribución muestral y *b*) mediante las fórmulas.

17.60 Resolver los problemas 17.58 y 17.59, pero esta vez con: *a*) dos letras *a* y dos letras *b*; *b*) una letra *a* y tres letras *b*, y *c*) una letra *a* y cuatro letras *b*.

17.61 Resolver los problemas 17.58 y 17.59, pero esta vez con dos letras *a* y cuatro letras *b*.

OTRAS APLICACIONES DE LA PRUEBA DE LAS RACHAS

17.62 Empleando como nivel de significancia 0.05, determinar si la muestra de las 40 calificaciones de la tabla 17.5 es aleatoria.

17.63 En la tabla 17.41 se da el precio de cierre de una acción en 25 días consecutivos. Al nivel de significancia 0.05, determinar si estos precios son aleatorios.

Tabla 17.41

10.375	11.125	10.875	10.625	11.500
11.625	11.250	11.375	10.750	11.000
10.875	10.750	11.500	11.250	12.125
11.875	11.375	11.875	11.125	11.750
11.375	12.125	11.750	11.500	12.250

17.64 Los primeros dígitos de $\sqrt{2}$ son 1.41421 35623 73095 0488$\cdots$. ¿Qué conclusión se puede sacar respecto a la aleatoriedad de estos dígitos?

17.65 ¿Qué conclusión se puede sacar respecto a la aleatoriedad de los dígitos siguientes?

a) $\sqrt{3} = 1.73205\ 08075\ 68877\ 2935\cdots$

b) $\pi = 3.14159\ 26535\ 89793\ 2643\cdots$

17.66 En el problema 17.62, mostrar que empleando la aproximación normal, el valor *p* es 0.105.

17.67 En el problema 17.63, mostrar que empleando la aproximación normal, el valor *p* es 0.168.

17.68 En el problema 17.64, mostrar que empleando la aproximación normal, el valor *p* es 0.485.

CORRELACIÓN DE RANGOS

17.69 En un concurso se pide a los jueces que ordenen a los candidatos (numerados del 1 al 8) de acuerdo con su preferencia. Los resultados se muestran en la tabla 17.42.

a) Encontrar el coeficiente de correlación de rangos.

b) Decidir si hay buena coincidencia entre los jueces.

Tabla 17.42

Primer juez	5	2	8	1	4	6	3	7
Segundo juez	4	5	7	3	2	8	1	6

17.70 La tabla 14.17, que se reproduce a continuación, da los índices de precios al consumidor, en Estados Unidos, para alimentos y atención médica desde 2000 hasta 2006, comparados con los precios de los años base, 1982 a 1984 (tomando la media como 100).

Año	2000	2001	2002	2003	2004	2005	2006
Alimentos	167.8	173.1	176.2	180.0	186.2	190.7	195.2
Medicina	260.8	272.8	285.6	297.1	310.1	323.2	336.2

Fuente: Bureau of Labor Statistics.

Según estos datos, encontrar la correlación de rangos de Spearman y el coeficiente de correlación de Pearson.

17.71 El coeficiente de correlación de rangos se obtiene empleando los datos ordenados por rangos en la fórmula del producto-momento del capítulo 14. Ilustrar esto empleando ambos métodos para resolver un problema.

17.72 Para datos agrupados, ¿puede obtenerse el coeficiente de correlación de rangos? Explicar e ilustrar la respuesta con un ejemplo.

CONTROL ESTADÍSTICO DE PROCESOS Y CAPACIDAD DE PROCESOS

18

ANÁLISIS GENERAL DE LAS GRÁFICAS DE CONTROL

Las variaciones que se presentan en cualquier proceso pueden deberse a *causas comunes* o a *causas especiales*. La variación natural de materiales, maquinaria y personas da lugar a las causas comunes de variación. Las causas especiales, también conocidas como *causas asignables*, se deben en la industria a desgaste excesivo de las herramientas, a un nuevo operador, a cambios en los materiales, a nuevos proveedores, etc. Uno de los propósitos de las *gráficas de control* es localizar, y si es posible, eliminar las causas especiales de variación. La estructura general de una gráfica de control consta de *límites de control* y de una *línea central*, como se muestra en la figura 18-1. Hay dos límites de control, el *límite superior de control* o UCL (por sus siglas en inglés) y el *límite inferior de control* o LCL (por sus siglas en inglés).

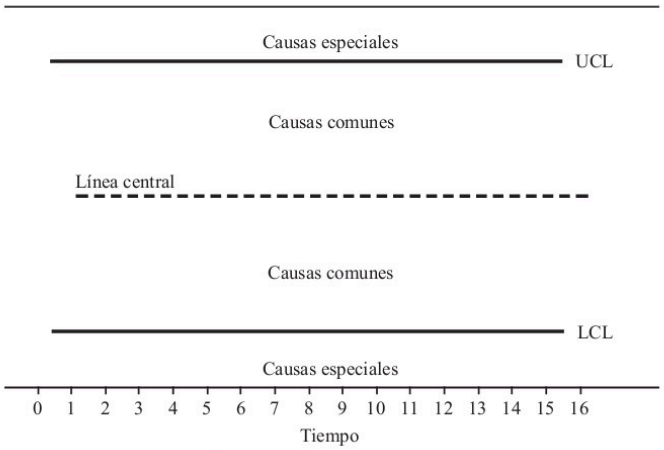


Figura 18-1 Las gráficas de control pueden ser de dos tipos (gráficas de control para variables y gráficas de control para atributos).

Cuando en una gráfica de control un punto cae fuera de los límites de control, se dice que el proceso está fuera de control estadístico. Además de los puntos fuera de control, hay otras anomalías que indican que un proceso está fuera de control. Éstas se verán más adelante. Lo deseable es que los procesos estén bajo control, de manera que su comportamiento sea previsible.

GRÁFICAS DE CONTROL DE VARIABLES Y GRÁFICAS DE CONTROL DE ATRIBUTOS

Las gráficas de control se pueden dividir en *gráficas de control de variables* y *gráficas de control de atributos*. Los términos "variable" y "atributo" se deben al tipo de datos que se recolectan del proceso. Si se miden características como tiempo, peso, volumen, longitud, caída de presión, concentración, etc., los datos obtenidos se consideran continuos y se conocen como *datos de variables*. Si se cuenta la cantidad de defectuosos en una muestra o la cantidad de defectos en determinado tipo de artículo, a los datos obtenidos se les llama *datos de atributos*. Se considera que los datos de variables son de nivel superior a los datos de atributos. En la tabla 18.1 se dan los nombres de algunas gráficas de control de variables y de control de atributos, así como los estadísticos que en ellas se grafican.

Tabla 18.1

Tipo de gráfica	Estadístico que se grafica
Gráfica *X*-barra y gráfica *R*	Promedios y rangos de subgrupos de datos de las variables
Gráfica *X*-barra y gráfica sigma	Promedios y desviaciones estándar de subgrupos de datos de las variables
Gráfica mediana	Mediana de subgrupos de datos de las variables
Gráficas de lecturas individuales	Mediciones individuales
Gráfica cusum	Suma acumulada de cada $\bar{X}$ menos el valor nominal
Gráfica de zonas	Pesos por zonas
Gráfica EWMA	Medias móviles con pesos exponenciales
Gráfica *P*	Proporción de artículos defectuosos en el total inspeccionado
Gráfica *NP*	Cantidad real de artículos defectuosos
Gráfica *C*	Cantidad de defectos por artículo en muestras de tamaño constante
Gráfica *U*	Cantidad de defectos por artículo en muestras de tamaño variable

En la tabla 18.1, las gráficas arriba de la línea punteada son gráficas de control de variables y las gráficas debajo de la línea punteada son gráficas de control de atributos. Actualmente, para la elaboración de gráficas suele emplearse algún software para estadística, como MINITAB.

GRÁFICAS *X*-BARRA Y GRÁFICAS *R*

Para entender la idea general de una gráfica *X*-barra considérese un proceso que tenga media μ y desviación estándar σ. Supóngase que el proceso se vigila tomando periódicamente muestras, a las que se les llama *subgrupos* de tamaño n, y calculando la media muestral $\bar{X}$ de cada una de ellas. El teorema del límite central asegura que la media de las medias muestrales es μ y la desviación estándar de las medias muestrales es $\sigma/\sqrt{n}$. La línea central de las medias muestrales se designa como μ y se considera que los límites inferior y superior de control están $3(\sigma/\sqrt{n})$ abajo y arriba de la línea central. El límite inferior de control está dado por la ecuación (*1*):

$$\text{LCL} = \mu - 3(\sigma/\sqrt{n}) \tag{1}$$

El límite superior de control está dado por la ecuación (*2*):

$$\text{UCL} = \mu + 3(\sigma/\sqrt{n}) \tag{2}$$

En un proceso distribuido normalmente, la media de un subgrupo caerá 99.7% de las veces entre los límites dados por (*1*) y (*2*). En la práctica no se conoce ni la media ni la desviación estándar del proceso y es necesario estimarlas. La media del proceso se estima mediante la media de las medias de las muestras periódicas. Esta media está dada por la ecuación (*3*), donde m es la cantidad de muestras de tamaño n tomadas periódicamente.

$$\bar{\bar{X}} = \frac{\sum \bar{X}}{m} \tag{3}$$

La media $\bar{\bar{X}}$, también puede encontrarse sumando todos los datos y dividiendo esta suma entre mn. La desviación estándar del proceso se estima promediando las desviaciones estándar o los rangos de los subgrupos, o bien usando un valor histórico de σ.

EJEMPLO 1 Se obtienen datos sobre la anchura de un producto. En 20 periodos se toman 5 observaciones de cada periodo. Los datos obtenidos se presentan en la tabla 18.2. El número de muestras periódicas es $m = 20$, el tamaño de la muestra o subgrupo es $n = 5$, la suma de todos los datos es 199.84 y la línea central es $\bar{\bar{X}} = 1.998$. La secuencia del menú de MINITAB "**Stat ⇒ Control charts ⇒ *X*bar**" se utilizó para procesar la gráfica de control que se muestra en la figura 18-2. Los datos de la tabla 18.2 se ingresan en una sola columna antes de aplicar la secuencia del menú anterior.

Tabla 18.2

1	2	3	4	5	6	7	8	9	10
2.000	2.007	1.987	1.989	1.997	1.983	1.966	2.004	2.009	1.991
1.988	1.988	1.983	1.989	2.018	1.972	1.982	1.998	1.994	1.989
1.975	2.002	2.006	1.997	1.999	2.002	1.995	2.011	2.020	2.000
1.994	1.978	2.019	1.976	1.990	1.991	2.020	1.991	2.000	2.016
1.991	2.012	2.021	2.007	2.003	1.997	2.008	1.972	2.006	2.037

11	12	13	14	15	16	17	18	19	20
2.004	1.988	1.996	1.999	2.018	1.986	2.002	1.988	2.011	1.998
1.980	1.991	2.005	1.984	2.009	2.010	1.969	2.031	1.976	2.003
1,998	2.003	1.996	1.988	2.023	2.012	2.018	1.978	1.998	2.016
1.994	1.997	2.008	2.011	2.010	2.013	1.984	1.987	2.023	1.996
2.006	1.985	2.007	2.005	1.993	1.988	1.990	1.990	1.998	2.009

La desviación estándar del producto puede estimarse de cuatro maneras distintas: sacando el promedio de los rangos de los 20 subgrupos, sacando el promedio de las desviaciones estándar de los 20 subgrupos, conjuntando las varianzas de los 20 subgrupos o mediante un valor histórico de σ, en caso de que se conozca alguno. MINITAB permite utilizar cualquiera de las cuatro opciones. En la figura 18-2 se grafican las 20 medias de las muestras que se presentan en la tabla 18.2. Esta gráfica indica que el proceso está bajo control. Las medias varían aleatoriamente respecto a la línea central y ninguna cae fuera de los límites de control.

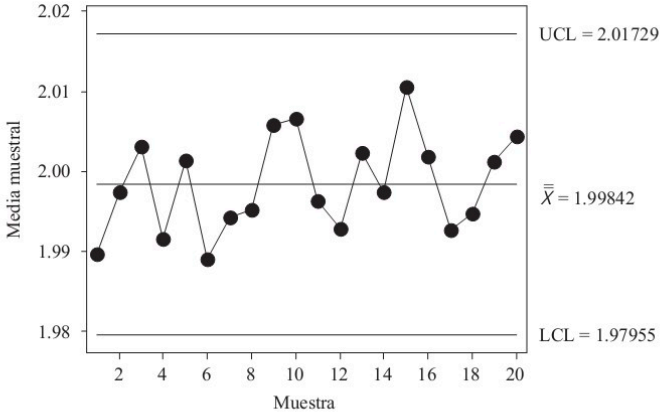


Figura 18-2 Gráfica *X*-barra para las anchuras.

Las *gráficas R* se usan para vigilar la variación del proceso. Para cada uno de los m subgrupos se calcula el rango R. La línea central de la gráfica R está dada por la ecuación (*4*)

$$\bar{R} = \frac{\sum R}{m} \tag{4}$$

Como en el caso de la gráfica *X*-barra, hay diversos métodos para estimar la desviación estándar del proceso.

EJEMPLO 2 Dados los datos de la tabla 18.2, el rango del primer subgrupo es $R_1 = 2.000 - 1.975 = 0.025$, el rango del segundo subgrupo es $R_2 = 2.012 - 1.978 = 0.034$. Los 20 rangos son: 0.025, 0.034, 0.038, 0.031, 0.028, 0.030, 0.054, 0.039, 0.026, 0.048, 0.026, 0.018, 0.012, 0.027, 0.030, 0.027, 0.049, 0.053, 0.047 y 0.020. La media de estos 20 rangos es 0.0327. En la figura 18-3 se presenta una gráfica de estos rangos elaborada con MINITAB. Esta gráfica R no muestra patrón inusual alguno respecto de la variabilidad. Para obtener la gráfica de control de la figura 18-3 se emplea la secuencia "**Stat → Control charts → *R* chart**" de MINITAB. Antes de aplicar esta secuencia, los datos de la tabla 18.2 deben ingresarse en una sola columna.

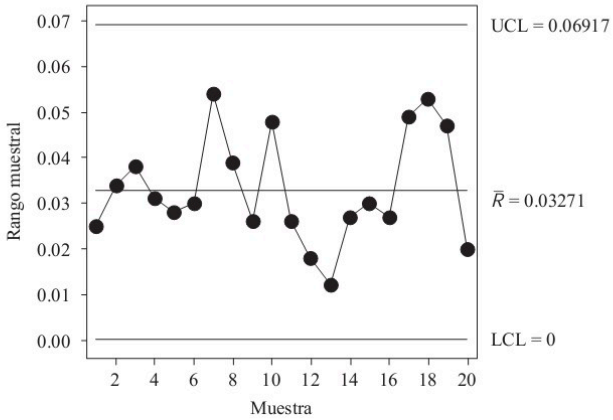


Figura 18-3 Gráfica *R* para las anchuras.

PRUEBAS PARA CAUSAS ESPECIALES

Además de un punto que caiga fuera de los límites de control, hay otros indicadores que sugieren falta de aleatoriedad en un proceso debida a causas especiales. En la tabla 18.3 se presentan ocho pruebas para causas especiales.

Tabla 18.3 Pruebas para causas especiales

1. Un punto a más de 3 sigmas de la línea central
2. Nueve puntos consecutivos de un mismo lado de la línea central
3. Seis puntos consecutivos, crecientes o decrecientes
4. Catorce puntos consecutivos, alternados arriba y abajo
5. Dos de tres puntos a más de 2 sigmas de la línea central (de un mismo lado)
6. Cuatro de cinco puntos a más de 1 sigma de la línea central (de un mismo lado)
7. Quince puntos consecutivos a más de 1 sigma de la línea central (de cualquier lado)
8. Ocho puntos consecutivos a más de 1 sigma de la línea central (de cualquier lado)

CAPACIDAD DE PROCESOS

Para llevar a cabo un análisis de la capacidad de un proceso, el proceso debe estar bajo control estadístico. Normalmente se supone que las características del proceso que van a ser medidas tienen distribución normal. Esto se puede comprobar empleando pruebas de normalidad, como la prueba de Kolmogorov-Smirnov, la prueba de Ryan-Joiner o la prueba de Anderson-Darling. La capacidad del proceso es una comparación entre el desempeño del proceso y los requerimientos del mismo. Los requerimientos del proceso determinan los *límites de especificación*. El LSL y el USL (por sus siglas en inglés) son, respectivamente, el *límite inferior de especificación* y el *límite superior de especificación*.

Los datos utilizados para determinar si un proceso está bajo control estadístico pueden emplearse para hacer el análisis de capacidad. A la distancia de 3 sigmas a ambos lados de la media se le conoce como *dispersión del proceso*. La media y la desviación estándar de las características del proceso suelen estimarse a partir de los datos obtenidos para el estudio del control estadístico del proceso.

EJEMPLO 3 Como se vio en el ejemplo 2, los datos de la tabla 18.2 provienen de un proceso que está bajo control estadístico. Se encuentra que la estimación de la media del proceso es 1.9984. Y la desviación estándar de las 100 observaciones es 0.013931. Supóngase que los límites de especificación son LSL = 1.970 y USL = 2.030. La prueba de Kolmogorov-Smirnov para normalidad se aplica usando MINITAB y se encuentra que no se puede rechazar la normalidad de la característica del proceso. Las *tasas de no conformes* se calculan como sigue. La proporción arriba del USL $= P(X > 2.030) = P[(X - 1.9984)/0.013931 > (2.030 - 1.9984)/0.013931] = P(Z > 2.27) = 0.0116$. Es decir, hay 0.0116(1 000 000) = 11 600 *partes por millón* (*ppm*) superiores al USL que son no conformes. Obsérvese que $P(Z > 2.27)$ puede encontrarse usando MINITAB, en lugar de buscar en las tablas de distribución normal estándar. Esto se hace como sigue. Se emplea la secuencia **Calc** → **Probability Distribution** → **Normal**.

Con $X = 2.27$ se obtiene

```
          x     P( X <= x )
    2.2700          0.9884
```

Se tiene $P(Z < 2.27) = 0.9884$, por lo tanto, $P(Z > 2.27) = 1 - 0.09884 = 0.0116$.

De igual manera, la proporción abajo del LSL $= P(X < 1.970) = P(Z < -2.04) = 0.0207$. Hay 20 700 ppm abajo del LSL que son no conformes. También aquí se emplea MINITAB para hallar el área bajo la curva normal estándar a la izquierda de −2.04.

La cantidad total de unidades no conformes es 11 600 + 20 700 = 32 300 ppm. Esto es, claramente, un número inaceptablemente elevado de unidades no conformes.

Supóngase que $\hat{\mu}$ es la estimación de la media de la característica del proceso y $\hat{\sigma}$ es la estimación de la desviación estándar de la característica del proceso, entonces la tasa de no conformes se estima como sigue. La proporción arriba del USL es igual a

$$P(X > \text{USL}) = P\left(Z > \frac{\text{USL} - \hat{\mu}}{\hat{\sigma}}\right)$$

y la proporción abajo del LSL es igual a

$$P(X < \text{LSL}) = P\left(Z < \frac{\text{LSL} - \hat{\mu}}{\hat{\sigma}}\right)$$

El *índice de capacidad del proceso* mide el potencial del proceso para satisfacer las especificaciones y se define como sigue:

$$C_{\text{P}} = \frac{\text{dispersión permitida}}{\text{dispersión medida}} = \frac{\text{USL} - \text{LSL}}{6\hat{\sigma}} \tag{5}$$

EJEMPLO 4 Dados los datos del proceso de la tabla 18.2, USL − LSL = 2.030 − 1.970 = 0.060, $6\hat{\sigma}$ = 6(0.013931) = 0.083586 y $C_{\text{p}} = 0.060/0.083586 = 0.72$.

El *índice* C_{PK} mide el desempeño del proceso y se define como sigue:

$$C_{\text{PK}} = \text{mínimo}\left\{\frac{\text{USL} - \hat{\mu}}{3\hat{\sigma}}, \quad \frac{\hat{\mu} - \text{LSL}}{3\hat{\sigma}}\right\} \tag{6}$$

EJEMPLO 5 Dados los datos del proceso del ejemplo 1,

$$C_{\text{PK}} = \text{mínimo}\left\{\frac{2.030 - 1.9984}{3(0.013931)}, \quad \frac{1.9984 - 1.970}{3(0.013931)}\right\} = \text{mínimo}\,\{0.76, 0.68\} = 0.68$$

En procesos que únicamente tienen límite inferior de especificación, el *índice inferior de capacidad* C_{PL} se define como sigue:

$$C_{\text{PL}} = \frac{\hat{\mu} - \text{LSL}}{3\hat{\sigma}} \tag{7}$$

En procesos que únicamente tienen límite superior de especificación, el *índice superior de capacidad* C_{PU} se define como sigue:

$$C_{\text{PU}} = \frac{\text{USL} - \hat{\mu}}{3\hat{\sigma}} \tag{8}$$

El C_{PK} se puede definir en términos del C_{PL} y del C_{PU} como sigue:

$$C_{\text{PK}} = \text{mín}\,\{C_{\text{PL}}, C_{\text{PU}}\} \tag{9}$$

La relación entre tasas de no conformes y C_{PL} y C_{PU} se obtiene como sigue:

$$P(X < \text{LSL}) = P\left(Z < \frac{\text{LSL} - \hat{\mu}}{\hat{\sigma}}\right) = P(Z < -3C_{\text{PL}}) \text{ dado que } -3C_{\text{PL}} = \frac{\text{LSL} - \hat{\mu}}{\hat{\sigma}}$$

$$P(X > \text{USL}) = P\left(Z > \frac{\text{USL} - \hat{\mu}}{\hat{\sigma}}\right) = P(Z > 3C_{\text{PU}}) \text{ dado que } 3C_{\text{PU}} = \frac{\text{USL} - \hat{\mu}}{\hat{\sigma}}$$

EJEMPLO 6 Supóngase que $C_{PL} = 1.1$, entonces la proporción de no conformes es $P(Z < -3(1.1)) = P(Z < -3.3)$. Esto se puede encontrar usando MINITAB, de la manera siguiente. Se da la secuencia "**Calc** ⇒ **Probability Distribution** ⇒ **Normal**".

El área acumulada a la izquierda de −3.3 está dada como:

Función de distribución acumulada

```
Normal with mean = 0 and standard deviation = 1
          x     P( X <= x )
       -3.3     0.00048348
```

Habrá 1 000 000 × 0.00048348 = 483 ppm de no conformes. Empleando esta técnica se puede elaborar una tabla en la que se relacione la tasa de no conformes con el índice de capacidad. Esto se da en la tabla 18.4.

Tabla 18.4

C_{PL} o C_{PU}	Proporción de no conformes	ppm
0.1	0.38208867	382089
0.2	0.27425308	274253
0.3	0.18406010	184060
0.4	0.11506974	115070
0.5	0.06680723	66807
0.6	0.03593027	35930
0.7	0.01786436	17864
0.8	0.00819753	8198
0.9	0.00346702	3467
1.0	0.00134997	1350
1.1	0.00048348	483
1.2	0.00015915	159
1.3	0.00004812	48
1.4	0.00001335	13
1.5	0.00000340	3
1.6	0.00000079	1
1.7	0.00000017	0
1.8	0.00000003	0
1.9	0.00000001	0
2.0	0.00000000	0

EJEMPLO 7 Usando la siguiente secuencia de MINITAB **Stat** ⇒ **Quality tools** ⇒ **Capability Analysis (Normal)** se obtiene un análisis de capacidad de los datos de la tabla 18.2. Estos resultados de MINITAB se muestran en la figura 18-4. En estos resultados se dan las tasas de no conformes, los índices de capacidad y algunas otras medidas. Las cantidades halladas en los ejemplos 3, 4 y 5 se acercan mucho a las medidas correspondientes mostradas en la figura. Las diferencias se deben a errores de redondeo, así como a diferentes métodos para estimar ciertos parámetros. La gráfica es muy ilustrativa, y señala la distribución de las mediciones muestrales en un histograma. La distribución poblacional de las mediciones del proceso aparece como una curva normal. Las áreas, a la derecha del USL y a la izquierda del LSL, en las colas bajo la curva normal, representan el porcentaje de productos no conformes. Multiplicando la suma de estos porcentajes por un millón, se obtiene la tasa, en ppm, de no conformes en el proceso.

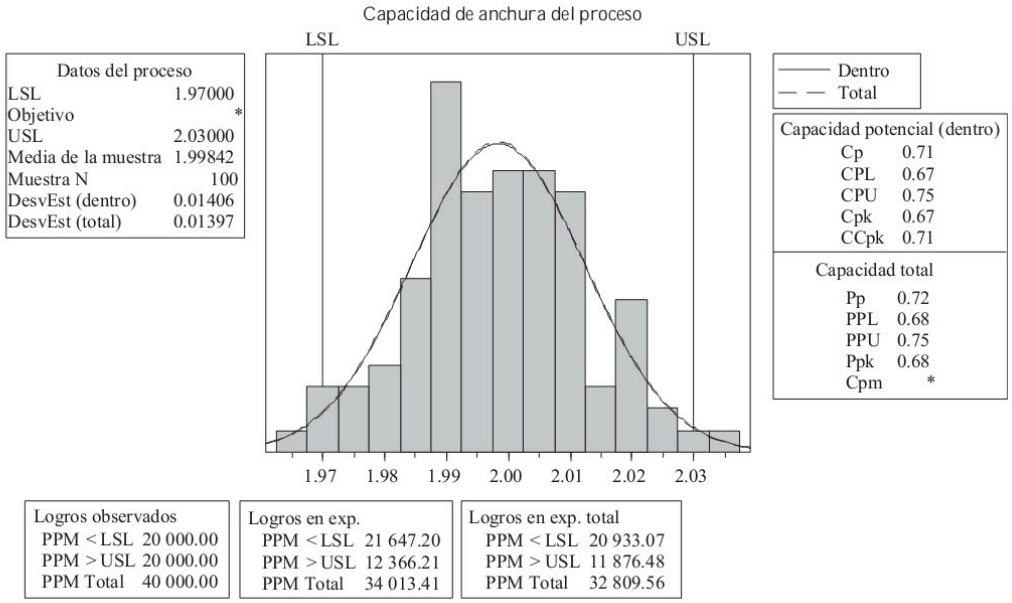


Figura 18-4 Varias medidas de la capacidad del proceso.

GRÁFICAS *P* Y *NP*

Cuando se categorizan o clasifican artículos producidos en masa, a los datos resultantes se les llama *datos de atributos*. Una vez establecidos los estándares que debe satisfacer un producto, se determinan las especificaciones. A un artículo que no satisface las especificaciones se le llama *artículo no conforme*. A un artículo que es no conforme y que no sirve para ser usado se le llama *artículo defectuoso*. Resulta más grave que un artículo sea defectuoso a que sea no conforme. Un artículo puede ser no conforme debido a un rayón o una decoloración, sin que sea un artículo defectuoso. Que un artículo no satisfaga una prueba de desempeño posiblemente hará que el producto sea clasificado como defectuoso y como no conforme. A los defectos encontrados en un artículo se les llama *no conformidades*. A las fallas irreparables se les llama *defectos*.

Para los datos de atributo se pueden emplear cuatro gráficas de control distintas. Estas cuatro gráficas son las gráficas *P*, las *NP*, las *C* y las *U*. Las gráficas *P* y las *NP* se basan en la distribución binomial y las gráficas *C* y *U* se basan en la distribución de Poisson. Las gráficas *P* se usan para vigilar la proporción de artículos no conformes producidos en un proceso. En el ejemplo 8 se ilustran las gráficas *P*, así como la notación que se emplea para describirlas.

EJEMPLO 8 Supóngase que cada 30 minutos se examinan 20 mascarillas para respiración y que por cada turno de 8 h se registra la cantidad de unidades defectuosas. La cantidad total examinada durante un turno es igual a $n = 20(16) = 320$. En la tabla 18.5 se dan los resultados obtenidos en 30 turnos. La línea central de la gráfica *P* corresponde a la proporción de defectuosos en los 30 turnos, y está dada por la cantidad total de defectuosos entre el total de examinados en los 30 turnos, es decir

$$\bar{p} = 72/9\,600 = 0.0075$$

La desviación estándar de la distribución binomial, correspondiente a esta gráfica, es

$$\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\frac{0.0075 \times 0.9925}{320}} = 0.004823$$

Los límites de control de 3 sigmas para este proceso son

Tabla 18.5

Turno #	Cantidad de defectuosos X_i	Proporción de defectuosos $P_i = X/n$	Turno #	Cantidad de defectuosos X_i	Proporción de defectuosos $P_i = X/n$
1	1	0.003125	16	2	0.006250
2	2	0.006250	17	0	0.000000
3	2	0.006250	18	4	0.012500
4	0	0.000000	19	1	0.003125
5	4	0.012500	20	7	0.021875
6	4	0.012500	21	4	0.012500
7	4	0.012500	22	1	0.003125
8	6	0.018750	23	0	0.000000
9	4	0.012500	24	4	0.012500
10	0	0.000000	25	4	0.012500
11	0	0.000000	26	3	0.009375
12	0	0.000000	27	2	0.006250
13	1	0.003125	28	0	0.000000
14	0	0.000000	29	0	0.000000
15	7	0.021875	30	5	0.015625

$$\bar{p} \pm 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{10}$$

El límite inferior de control es LCL $= 0.0075 - 3(0.004823) = -0.006969$. Cuando el LCL es negativo, se considera igual a cero, ya que la proporción de defectuosos en una muestra no es posible que sea negativa. El límite superior de control es UCL $= 0.0075 + 3(0.004823) = 0.021969$.

La gráfica P de este proceso, empleando MINITAB, se obtiene con la secuencia **Stat** → **Control charts** → **P**. En la figura 18-5 se muestra la gráfica P. Aunque al parecer las muestras 15 y 20 indican la presencia de una causa especial, cuando se compara la proporción de defectuosos en estas muestras 15 y 20 (ambas igual a 0.021875) con el UCL $= 0.021969$, se ve que estos puntos no están más allá del UCL.

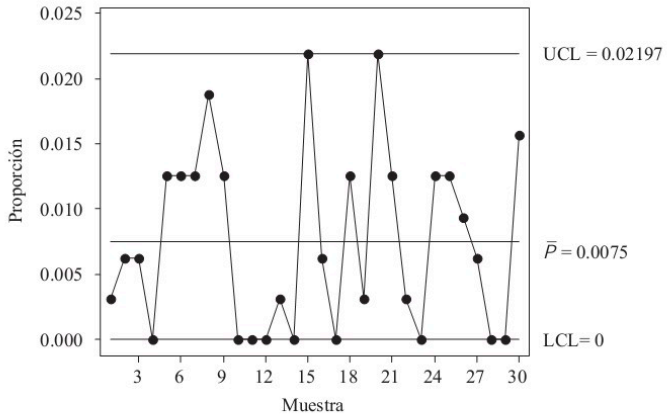


Figura 18-5 Con la gráfica p se vigila la proporción de defectuosos.

Con las *gráficas NP* se vigila el número de defectuosos, en lugar de la proporción de defectuosos. La gráfica *NP* es preferida por muchos debido a que es más fácil de entender que la proporción de defectuosos, tanto para los técnicos de calidad como para los operadores. La línea central en la gráfica *NP* está dada por $n\bar{p}$ y los límites de control de 3 sigmas son

$$n\bar{p} \pm 3\sqrt{n\bar{p}(1-\bar{p})} \tag{11}$$

EJEMPLO 9 Dados los datos de la tabla 18.5, la línea central está dada por $n\bar{p} = 320(.0075) = 2.4$ y los límites de control son $\text{LCL} = 2.4 - 4.63 = -2.23$, que se toma igual a 0 y $\text{UCL} = 2.4 + 4.63 = 7.03$. Si en un turno se encuentran 8 o más defectuosos, el proceso estará fuera de control. Con MINITAB, la solución se encuentra usando la secuencia

Stat → Control charts → NP

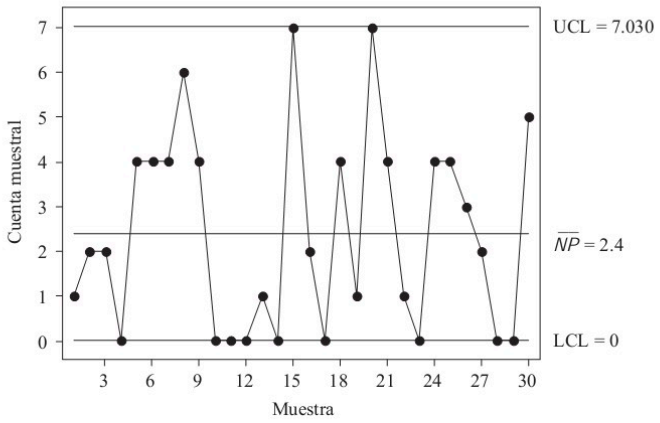


Figura 18-6 Con la gráfica NP se vigila el número de defectuosos.

Antes de ejecutar esta secuencia, el número de defectuosos por muestra debe ser ingresado en alguna columna de la hoja de cálculo. En la figura 18-6 se muestra una gráfica NP.

OTRAS GRÁFICAS DE CONTROL

Este capítulo es sólo una introducción al uso de las gráficas de control como ayuda en el proceso de control estadístico. En la tabla 18.1 se enumeran varias de las muchas gráficas de control que se emplean actualmente en la industria. Para facilitar los cálculos, en las plantas de producción suele usarse la *gráfica mediana*. En lugar de las medias de las muestras se grafican las medianas de las muestras. Si el tamaño de la muestra es non, entonces la mediana es simplemente el valor de en medio en el conjunto de valores ordenados de menor a mayor.

Cuando el volumen de producción es pequeño suelen emplearse las *gráficas de lecturas individuales*. En este caso, el subgrupo o muestra consta de una sola observación. A las gráficas de lecturas individuales también se les llama gráficas *X*.

Una *gráfica de zonas* está dividida en cuatro zonas. La zona 1 son los valores a no más de 1 desviación estándar de la media, la zona 2 son los valores entre 1 y 2 desviaciones estándar de la media, la zona 3 son los valores entre 2 y 3 desviaciones estándar de la media, y la zona 4 son los valores a 3 o más desviaciones estándar de la media. A las cuatro zonas se les asignan pesos. Los pesos de los puntos a un mismo lado de la línea central se suman y si la suma acumulada es mayor o igual al peso asignado a la zona 4, esto se considera como una señal de que el proceso está fuera de control. La suma acumulada se hace igual a cero después de que el proceso se ha considerado fuera de control, o cuando el siguiente punto graficado cruza la línea central.

Las gráficas de medias móviles con pesos exponenciales (*gráfica EWMA*, por sus siglas en inglés) son una alternativa a las gráficas de lecturas individuales o a las gráficas *X*-barra y proporcionan una rápida respuesta a cualquier desplazamiento del promedio del proceso. En las gráficas EWMA se incorpora información de todos los subgrupos anteriores, no sólo del subgrupo presente.

Las sumas acumuladas de las desviaciones del valor objetivo del proceso se utilizan en las *gráficas cusum*. Tanto las gráficas EWMA como las gráficas cusum permiten detectar fácilmente cualquier desplazamiento del proceso.

Cuando lo que interesa es la cantidad de no conformidades o de defectos en un producto y no simplemente determinar si el producto está defectuoso o no, se usan las *gráficas C* o las *gráficas U*. Cuando se usan estas gráficas es importante definir una *unidad de inspección*. La unidad de inspección se define como la unidad de producción a ser muestreada y examinada respecto de no conformidades. Si sólo hay una unidad de inspección por muestra, se usa una gráfica *C*, y si el número de unidades de inspección por muestra varía, se usa una gráfica *U*.

PROBLEMAS RESUELTOS

GRÁFICAS *X*-BARRA Y GRÁFICAS *R*

18.1 En un proceso industrial se llenan paquetes de avena para desayuno. La media de llenado de este proceso es 510 gramos (g) y la desviación estándar del llenado es 5 g. Cada hora se toman cuatro paquetes y el peso medio del subgrupo de cuatro pesos se emplea para vigilar el proceso respecto a causas especiales y para ayudar a mantener el proceso bajo control estadístico. Hallar los límites inferior y superior de control de la gráfica de control *X*-barra.

SOLUCIÓN

En este problema se supone que se conocen μ y σ y que son iguales a 510 y 5, respectivamente. Cuando no se conocen ni μ ni σ, será necesario estimarlas. El límite inferior de control es $\text{LCL} = \mu - 3(\sigma/\sqrt{n}) = 510 - 3(2.5) = 502.5$ y el límite superior de control es $\text{UCL} = \mu + 3(\sigma/\sqrt{n}) = 510 + 3(2.5) = 517.5$.

Tabla 18.6

Periodo 1	Periodo 2	Periodo 3	Periodo 4	Periodo 5	Periodo 6	Periodo 7	Periodo 8	Periodo 9	Periodo 10
2.000	2.007	1.987	1.989	1.997	1.983	1.966	2.004	2.009	1.991
1.988	1.988	1.983	1.989	2.018	1.972	1.982	1.998	1.994	1.989
1.975	2.002	2.006	1.997	1.999	2.002	1.995	2.011	2.020	2.000
1.994	1.978	2.019	1.976	1.990	1.991	2.020	1.991	2.000	2.016
1.991	2.012	2.021	2.007	2.003	1.997	2.008	1.972	2.006	2.037

Periodo 11	Periodo 12	Periodo 13	Periodo 14	Periodo 15	Periodo 16	Periodo 17	Periodo 18	Periodo 19	Periodo 20
2.004	1.988	1.996	1.999	2.018	2.025	2.002	1.988	2.011	1.998
1.980	1.991	2.005	1.984	2.009	2.022	1.969	2.031	1.976	2.003
1.998	2.003	1.996	1.988	2.023	2.035	2.018	1.978	1.998	2.016
1.994	1.997	2.008	2.011	2.010	2.013	1.984	1.987	2.023	1.996
2.006	1.985	2.007	2.005	1.993	2.020	1.990	1.990	1.998	2.009

18.2 La tabla 18.6 contiene las anchuras de un producto obtenidas en 20 periodos. Los límites de control para la gráfica X-barra son LCL = 1.981 y UCL = 2.018. ¿Está alguna de las medias de los grupos fuera de estos límites de control?

SOLUCIÓN

Las medias de los 20 subgrupos son 1.9896, 1.9974, 2.0032, 1.9916, 2.0014, 1.9890, 1.9942, 1.9952, 2.0058, 2.0066, 1.9964, 1.9928, 2.0024, 1.9974, 2.0106, **2.0230**, 1.9926, 1.9948, 2.0012 y 2.0044, respectivamente. La media número dieciséis, 2.0230, está fuera del límite superior de control. Todas las demás medias están dentro de los límites de control.

18.3 Volviendo al problema 18.2, se encuentra que precisamente antes de muestrear el grupo número dieciséis hubo una falla. Este subgrupo se elimina, se vuelven a calcular los límites de control y se encuentra que los nuevos límites de control son LCL = 1.979 y UCL = 2.017. ¿Hay alguna otra media, además de la media del subgrupo dieciséis, que esté fuera de los nuevos límites de control?

SOLUCIÓN

Ninguna de las medias dadas en el problema 18.2, además de la número dieciséis, cae fuera de los nuevos límites. Suponiendo que la nueva gráfica satisfaga todas las otras pruebas para causas especiales, dadas en la tabla 18.3, los límites de control dados en este problema pueden emplearse para vigilar el proceso.

18.4 Verificar los límites de control dados en el problema 18.2. Estimar la desviación estándar del proceso conjuntando las 20 varianzas muestrales.

SOLUCIÓN

La media de las 100 observaciones muestrales es 1.999. Una manera de hallar la varianza conjunta de estas 20 muestras es tratar estas 20 muestras, cada una con cinco observaciones, como una clasificación en un sentido. El error cuadrado medio dentro de los tratamientos es igual a la varianza conjunta de las 20 muestras. Empleando el análisis de MINITAB para un diseño en un sentido, se obtiene la siguiente tabla de análisis de varianza

```
Analysis of Variance
Source   DF        SS        MS      F       P
Factor   19  0.006342  0.000334   1.75   0.044
Error    80  0.015245  0.000191
Total    99  0.021587
```

La estimación de la desviación estándar es $\sqrt{0.000191} = 0.01382$. El límite inferior de control es LCL $= 1.999 - 3(0.01382/\sqrt{5}) = 1.981$ y el límite superior de control es UCL $= 1.999 + 3(0.01382/\sqrt{5}) = 2.018$.

PRUEBAS PARA CAUSAS ESPECIALES

18.5 En la tabla 18.7 se presentan los datos de 20 subgrupos, cada uno de tamaño 5. La gráfica X-barra se da en la figura 18-7. ¿Qué efectos tuvo en el proceso un cambio a un nuevo proveedor en el periodo 10? ¿Cuál es la prueba para causas especiales, de la tabla 18.3, que no satisface el proceso?

SOLUCIÓN

En la gráfica de control de la figura 18-7 se observa que el cambio al nuevo proveedor ocasionó un aumento en la anchura. Este cambio después del periodo 10 es evidente. El 6 que aparece en la gráfica de la figura 18-7 indica que no se satisface la prueba 6 de la tabla 18.3. Cuatro de cinco puntos están a más de 1 sigma de la línea central (del mismo lado). Los cinco puntos corresponden a los subgrupos 4 a 8.

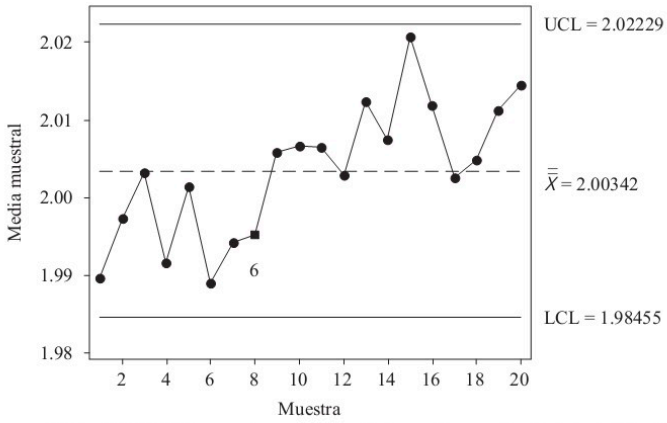


Figura 18-7 MINITAB, puntos que no satisfacen la prueba 6 de la tabla 18.3.

Tabla 18.7

Periodo 1	Periodo 2	Periodo 3	Periodo 4	Periodo 5	Periodo 6	Periodo 7	Periodo 8	Periodo 9	Periodo 10
2.000	2.007	1.987	1.989	1.997	1.983	1.966	2.004	2.009	1.991
1.988	1.988	1.983	1.989	2.018	1.972	1.982	1.998	1.994	1.989
1.975	2.002	2.006	1.997	1.999	2.002	1.995	2.011	2.020	2.000
1.994	1.978	2.019	1.976	1.990	1.991	2.020	1.991	2.000	2.016
1.991	2.012	2.021	2.007	2.003	1.997	2.008	1.972	2.006	2.037

Periodo 11	Periodo 12	Periodo 13	Periodo 14	Periodo 15	Periodo 16	Periodo 17	Periodo 18	Periodo 19	Periodo 20
2.014	1.998	2.006	2.009	2.028	1.996	2.012	1.998	2.021	2.008
1.990	2.001	2.015	1.994	2.019	2.020	1.979	2.041	1.986	2.013
2.008	2.013	2.006	1.998	2.033	2.022	2.028	1.988	2.008	2.026
2.004	2.007	2.018	2.021	2.020	2.023	1.994	1.997	2.033	2.006
2.016	1.995	2.017	2.015	2.003	1.998	2.000	2.000	2.008	2.019

CAPACIDAD DEL PROCESO

18.6 Volver al problema 18.2. Después de determinar una causa especial relacionada con el subgrupo 16, se elimina ese subgrupo. La anchura media se estima hallando la media de los datos de los 19 subgrupos restantes. Y la desviación estándar se estima hallando la desviación estándar de estos mismos datos. Si los límites de especificación son LSL = 1.960 y USL = 2.040, hallar el índice inferior de capacidad, el índice superior de capacidad y el índice C_{PK}.

SOLUCIÓN

Empleando las 95 mediciones restantes, después de excluir al subgrupo 16, se encuentra que $\hat{\mu} = 1.9982$ y $\hat{\sigma} = 0.01400$. El índice inferior de capacidad es

$$C_{\text{PL}} = \frac{\hat{\mu} - \text{LSL}}{3\hat{\sigma}} = \frac{1.9982 - 1.960}{0.0420} = 0.910$$

el índice superior de capacidad es

$$C_{\text{PU}} = \frac{\text{USL} - \hat{\mu}}{3\hat{\sigma}} = \frac{2.040 - 1.9982}{0.042} = 0.995$$

y $C_{\text{PK}} = \text{mín}\{C_{\text{PL}}, C_{\text{PU}}\} = 0.91$.

18.7 Volver al problema 18.1. *a*) Encontrar el porcentaje de no conformes si LSL = 495 y USL = 525. *b*) Encontrar las ppm de no conformes si LSL = 490 y USL = 530.

SOLUCIÓN

a) Suponiendo que el llenado tenga una distribución normal, el área bajo la curva normal abajo del LSL se encuentra empleando el comando de EXCEL `=NORMDIST(495,510,5,1)`, con el que se obtiene 0.001350. Por simetría, el área bajo la curva normal arriba del USL también es 0.001350. El total del área fuera de los límites de especificación es 0.002700. Las ppm de no conformes son 0.002700(1 000 000) = 2 700.

b) El área bajo la curva normal correspondiente a LSL = 490 y a USL = 530 se halla de manera similar y es 0.000032 + 0.000032 = 0.000064. Se encuentra que las partes por millón son 0.000064(1 000 000) = 64.

GRÁFICAS *P* Y *NP*

18.8 Se inspeccionan circuitos impresos para detectar soldaduras imperfectas. A lo largo de 30 días, diariamente se inspeccionan 500 circuitos impresos. En la tabla 18.8 se presentan las cantidades de defectuosos. Elaborar una gráfica *P* y localizar las causas especiales.

Tabla 18.8

Día	1	2	3	4	5	6	7	8	9	10
# Defectuosos	2	0	2	5	2	4	5	1	2	3
Día	11	12	13	14	15	16	17	18	19	20
# Defectuosos	3	2	0	4	3	8	10	4	4	5
Día	21	22	23	24	25	26	27	28	29	30
# Defectuosos	2	4	3	2	3	3	2	1	1	2

SOLUCIÓN

Los límites de confianza son

$$\bar{p} = 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

La línea central es $\bar{p} = 92/15\,000 = 0.00613$ y la desviación estándar es

$$\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = \sqrt{\frac{(0.00613)(0.99387)}{500}} = 0.00349$$

El límite inferior de control es $0.00613 - 0.01047 = -0.00434$, que se toma igual a 0, ya que las proporciones no pueden ser negativas. El límite superior de control es $0.00613 + 0.01047 = 0.0166$. El día 17 la proporción de defectuosos es $P_{17} = 10/500 = 0.02$; es el único día en que la proporción es mayor al límite superior.

18.9 Dados los datos del problema 18.8, proporcionar los límites de control de la gráfica-*NP*.

SOLUCIÓN

Los límites de control para el número de defectuosos son $n\bar{p} = 3\sqrt{n\bar{p}(1-\bar{p})}$. La línea central es $n\bar{p} = 3.067$. El límite inferior es 0 y el límite superior es 8.304.

18.10 Supóngase que se empacan mascarillas para respiración en cajas con 25 o con 50 unidades. Durante cada turno, a intervalos de 30 minutos, se toma de manera aleatoria una caja y se determina la cantidad de defectuosas en ella. La caja puede contener 25 o 50 mascarillas. La cantidad de mascarillas examinadas por turno varía entre 400 y 800. En la tabla 18.9 se presentan los datos. Usar MINITAB para elaborar la gráfica de control para la proporción de defectuosas.

Tabla 18.9

Turno #	Tamaño de la muestra n_i	Cantidad de defectuosas X_i	Proporción de defectuosas $P_i = X_i/n_i$
1	400	3	0.0075
2	575	7	0.0122
3	400	1	0.0025
4	800	7	0.0088
5	475	2	0.0042
6	575	0	0.0000
7	400	8	0.0200
8	625	1	0.0016
9	775	10	0.0129
10	425	8	0.0188
11	400	7	0.0175
12	400	3	0.0075
13	625	6	0.0096
14	800	5	0.0063
15	800	4	0.0050
16	800	7	0.0088
17	475	9	0.0189
18	800	9	0.0113
18	750	9	0.0120
20	475	2	0.0042

SOLUCIÓN

Cuando varía el tamaño de la muestra, la línea central siempre es la misma, es decir, es la proporción de defectuosos en todas las muestras. Pero la desviación estándar varía de una muestra a otra y los límites de control que se obtienen son límites de control escalonados. Los límites de control son

$$\bar{p} = 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}}$$

La línea central es $\bar{p} = 108/11\,775 = 0.009172$. Para el primer subgrupo, se tiene $n_i = 400$

$$\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}} = \sqrt{\frac{(0.009172)(0.990828)}{400}} = 0.004767$$

y $3(0.004767) = 0.014301$. El límite inferior del subgrupo 1 es 0 y el límite superior es $0.009172 + 0.014301 = 0.023473$. Los límites de los turnos restantes se determinan de igual manera. Estos límites cambiantes dan lugar a límites superiores de control escalonados, como se muestran en la figura 18-8.

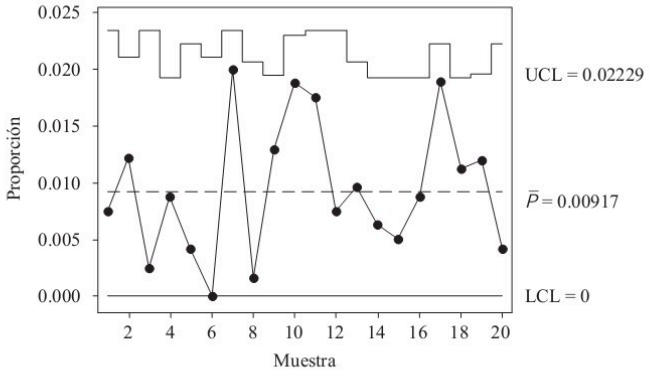


Figura 18-8 Gráfica *P* para tamaños de muestra variables.

OTRAS GRÁFICAS DE CONTROL

18.11 En los casos en que las mediciones resultan muy costosas, los datos se obtienen lentamente, o cuando la producción es bastante homogénea lo indicado es una *gráfica de lecturas o mediciones individuales de rangos móviles*. Los datos consisten en una sola medición tomada en diferentes momentos. La línea central es la media de todas las mediciones individuales y la variación se estima mediante el uso de *rangos móviles*. Normalmente, los rangos móviles se calculan restando los valores de datos adyacentes y tomando el valor absoluto del valor resultante. En la tabla 18.10 se presentan las mediciones codificadas de la resistencia a la ruptura de un costoso cable empleado en los aviones. Del proceso de producción, se toma un cable por día y se prueba. Proporcionar la gráfica de lecturas individuales generada con MINITAB e interpretar los resultados.

Tabla 18.10

Día	1	2	3	4	5	6	7	8	9	10
Resistencia	491.5	502.0	505.5	499.6	504.1	501.3	503.5	504.3	498.5	508.8
Día	11	12	13	14	15	16	17	18	19	20
Resistencia	515.4	508.0	506.0	510.9	507.6	519.1	506.9	510.9	503.9	507.4

SOLUCIÓN

Se emplea la secuencia siguiente **Stats** → **Control charts** → **individuals**.

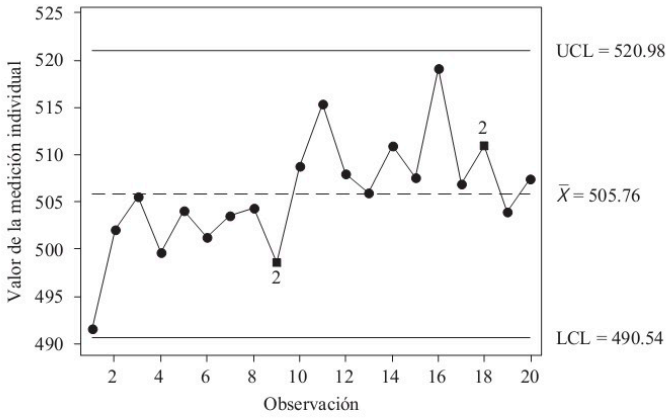


Figura 18-9 Gráfica de mediciones individuales para la resistencia.

En la figura 18-9 se presenta la gráfica de lecturas individuales correspondiente a los datos de la tabla 18.10. En esta gráfica de control se grafican los valores de las lecturas individuales de la tabla 18.10. El 2 que aparece en las semanas 9 y 18 de la gráfica de control hace referencia a la segunda prueba para causas especiales de la tabla 18.3. Esta indicación de una causa especial corresponde a nueve puntos consecutivos de un mismo lado de la línea central. En el periodo 10, un aumento de la temperatura del proceso ocasionó incremento en la resistencia a la ruptura. Este cambio en la resistencia a la ruptura resultó en puntos debajo de la línea central antes del periodo 10 y puntos sobre la línea central después del periodo 10.

18.12 La *gráfica EWMA de promedios móviles exponencialmente ponderados* se usa para detectar pequeños cambios respecto a un valor objetivo *t*. Los puntos de la gráfica EWMA están dados por la ecuación siguiente:

$$\hat{x}_i = w\bar{x}_i + (1 - w)\hat{x}_{i-1}$$

Para ilustrar el uso de esta ecuación, supóngase que los datos de la tabla 18.7 hayan sido seleccionados de un proceso cuyo valor objetivo sea 2 000. El valor inicial $\hat{x}_0$ se elige igual al valor objetivo, 2 000. Como peso w suele elegirse un valor entre 0.10 y 0.30. Si no se especifica ningún valor, MINITAB utiliza 0.20. El primer punto de la gráfica EWMA será $\hat{x}_1 = w\bar{x}_1 + (1 - w)\,\hat{x}_0 + 0.20(1.9896) + 0.80(2.000) = 1.9979$. El segundo punto de la gráfica será $\hat{x}_2 = w\bar{x}_2 + (1 - w)\,\hat{x}_1 = 0.20(1.9974) + 0.80(1.9979) = 1.9978$, etc. El análisis de MINITAB se obtiene empleando la secuencia siguiente **Stat** → **Control charts** → **EWMA**. Es necesario proporcionar a MINITAB el valor objetivo. En la figura 18-10 se muestran los resultados. De acuerdo con la figura, determinar en qué subgrupo se desvía el proceso del valor objetivo.

SOLUCIÓN

La gráfica de los valores $\hat{x}_i$ cruza el límite superior de control con el punto correspondiente al periodo 15. Éste es el punto en el que se concluirá que el proceso se aleja del valor objetivo. Obsérvese que la gráfica EWMA tiene límites de control escalonados.

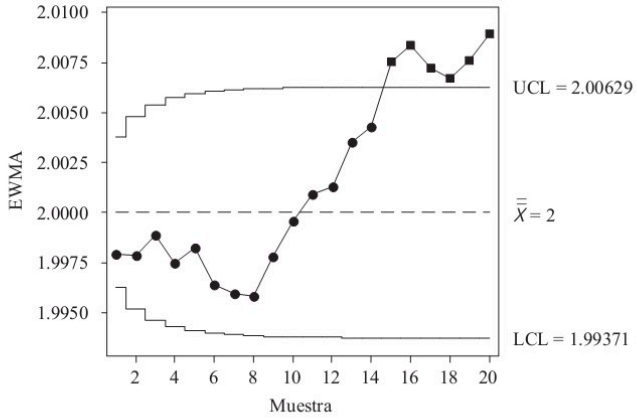


Figura 18-10 Gráfica de promedios móviles exponencialmente ponderados.

18.13 Una *gráfica de zonas* se divide en cuatro zonas. La zona 1 se define como los valores a no más de 1 desviación estándar de la media, la zona 2 se define como los valores entre 1 y 2 desviaciones estándar de la media, la zona 3 se define como los valores entre 2 y 3 desviaciones estándar de la media, y la zona 4 como los valores a 3 o más desviaciones estándar de la media. Si no se especifica otra cosa, MINITAB asigna a las zonas 1 a 4 los valores 0, 2, 4 y 8, respectivamente. Los puntos que se encuentran de un mismo lado de la línea central se suman. Si una suma acumulada es mayor o igual al peso asignado a la zona 4, eso se considera como una señal de que el proceso está fuera de control. Después de que un proceso se señala como fuera de control o cuando el siguiente punto graficado cruza la línea central, la suma acumulada se iguala a 0. En la figura 18-11 se presenta el análisis de MINITAB empleando una gráfica de zonas para los datos de la tabla 18.6. La secuencia para obtener esta gráfica es **Stat** → **Control charts** → **Zone**. ¿Qué puntos se encuentran fuera de control en esta gráfica de zona?

SOLUCIÓN

El subgrupo 16 corresponde a un punto fuera de control. La puntuación de zona correspondiente al subgrupo 16 es 10, que es mayor a la puntuación asignada a la zona 4, con lo que en el proceso se localiza un periodo fuera de control.

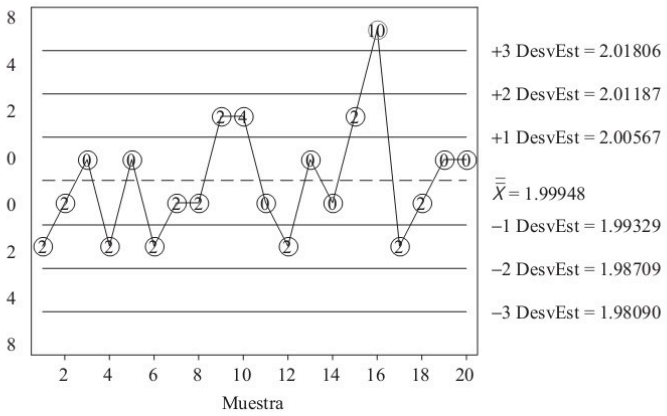


Figura 18-11 Gráfica de zonas para las anchuras.

18.14 Cuando lo que interesa es el número de no conformidades o de defectos en un producto, y no sólo determinar si el producto está o no defectuoso, se usa la *gráfica C* o la *gráfica U*. Para usar estas gráficas es importante definir la *unidad de inspección*. La unidad de inspección se define como la unidad de producción (de salida) a ser muestreada y examinada respecto a no conformidades. Si sólo hay una unidad de inspección por muestra, se usa una gráfica *C*; si la cantidad de unidades de inspección por muestra varía, se usa una gráfica *U*.

La fabricación de productos en rollo, como papel, películas, textiles, plásticos, etc., es un área en la que se usan la gráfica *C* y la gráfica *U*. No conformidades o defectos, como la aparición de puntos negros en una película fotográfica, atadijos de fibras, manchas, agujeritos, marcas por electricidad estática, suelen presentarse en algún grado en la fabricación de productos en rollo. El propósito de las gráficas *C* y *U* es garantizar que en el resultado del proceso la ocurrencia de tales inconformidades permanezca dentro de un nivel aceptable. Estas no conformidades suelen presentarse en forma aleatoria e independiente unas de otras en toda el área del producto. En estos casos, para elaborar la gráfica de control se emplea la distribución de Poisson. La línea central de una gráfica *C* se localiza en $\bar{c}$, la cantidad media de no conformidades en todos los subgrupos. La desviación estándar en la distribución de Poisson es $\sqrt{\bar{c}}$, y por lo tanto los límites de control 3 sigma son $\bar{c} \pm 3\sqrt{\bar{c}}$. Es decir, el límite inferior de control es LCL $= \bar{c} - 3\sqrt{\bar{c}}$ y el límite superior de control es UCL $= \bar{c} + 3\sqrt{\bar{c}}$.

Cuando se aplica un recubrimiento a un material, suelen formarse pequeñas no conformidades llamadas aglomerados. En los rollos jumbo de un producto se registra la cantidad de aglomerados por 5 pies (ft) de rollo. En la tabla 18.11 se presentan los resultados en 24 de estos rollos. ¿Hay algún punto fuera de los límites de control 3 sigma?

Tabla 18.11

Rollo jumbo #	1	2	3	4	5	6	7	8	9	10	11	12
Aglomerados	3	3	6	0	7	5	3	6	3	5	2	2
Rollo jumbo #	13	14	15	16	17	18	19	20	21	22	23	24
Aglomerados	2	7	6	4	7	8	5	13	7	3	3	7

SOLUCIÓN

La cantidad media de aglomerados por rollo jumbo es igual a la cantidad total de aglomerados dividida entre 24, esto es, $\bar{c} = 117/24 = 4.875$. La desviación estándar es $\sqrt{\bar{c}} = 2.208$. El límite inferior de control es LCL $= 4.875 - 3(2.208) = -1.749$. Como este valor es negativo, se toma 0 como límite inferior. El límite superior es UCL $= 4.875 + 3(2.208) = 11.499$. En el rollo jumbo # 20 hay una condición fuera de control, ya que la cantidad de aglomerados, 13, es mayor al límite superior de control, 11.499.

18.15 Este problema es continuación del problema 18.14. Antes de hacer este problema se deberá revisar el problema 18.14. En la tabla 18.12 se dan los datos de 20 rollos jumbo. En la tabla se da el número del rollo, la longitud del rollo inspeccionada para detectar aglomerados, la cantidad de unidades inspeccionadas (recuérdese que según el problema 18.14, una unidad de inspección es 5 ft), la cantidad de aglomerados encontrados en la longitud inspeccionada y la cantidad de aglomerados por unidad inspeccionada. La línea central de la gráfica *U* es $\bar{u}$, la suma de la columna 4 dividida entre la suma de la columna 3. Sin embargo, la desviación estándar cambia de una muestra a otra y hace que los límites de control sean límites de control escalonados. El límite inferior de control de la muestra *i* es UCL $= \bar{u} - 3\sqrt{\bar{u}/n_i}$ y el límite superior de control de la muestra *i* es UCL $= \bar{u} + 3\sqrt{\bar{u}/n_i}$.

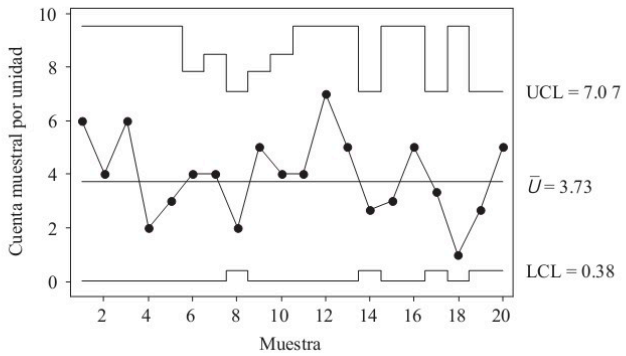
Tabla 18.12

Rollo jumbo #	Longitud inspeccionada	# de unidades inspeccionadas, n_i	# de aglomerados	$u_i =$ Col. 4/Col. 3
1	5.0	1.0	6	6.00
2	5.0	1.0	4	4.00
3	5.0	1.0	6	6.00
4	5.0	1.0	2	2.00
5	5.0	1.0	3	3.00
6	10.0	2.0	8	4.00
7	7.5	1.5	6	4.00
8	15.0	3.0	6	2.00
9	10.0	2.0	10	5.00
10	7.5	1.5	6	4.00
11	5.0	1.0	4	4.00
12	5.0	1.0	7	7.00
13	5.0	1.0	5	5.00
14	15.0	3.0	8	2.67
15	5.0	1.0	3	3.00
16	5.0	1.0	5	5.00
17	15.0	3.0	10	3.33
18	5.0	1.0	1	1.00
19	15.0	3.0	8	2.67
20	15.0	3.0	15	5.00

Usar MINITAB para elaborar la gráfica de control para este problema y determinar si el proceso está bajo control.

SOLUCIÓN

La línea central de la gráfica U es $\bar{u}$, la suma de la columna 4 entre la suma de la columna 3. Pero la desviación estándar cambia de una muestra a otra y hace que los límites de control sean escalonados. El límite inferior de control de la muestra i es LCL $= \bar{u} - 3\sqrt{\bar{u}/n_i}$ y el límite superior de control de la muestra i es UCL $= \bar{u} + 3\sqrt{\bar{u}/n_i}$. La línea central correspon-



Prueba realizada con tamaños de muestra diferentes

Figura 18-12 Gráfica U para aglomerados.

diente a los datos anteriores es $\bar{u} = 123/33 = 3.73$. La solución de MINITAB se obtiene mediante la secuencia **Stat** → **Control Charts** → **U**.

La información que requiere MINITAB para elaborar la gráfica *U* es la dada en las columnas 3 y 4 de la tabla 18.12. En la figura 18-12 se muestra la gráfica *U* de los datos de la tabla 18.12. La gráfica de control no indica que ningún periodo esté fuera de control.

PROBLEMAS SUPLEMENTARIOS

GRÁFICA *X*-BARRA Y GRÁFICA *R*

18.16 En la tabla 18.13 se presentan los datos de 10 subgrupos, cada uno de tamaño 4. Para cada subgrupo calcular $\bar{X}$ y R, así como $\bar{\bar{X}}$ y $\bar{R}$. En una gráfica señalar los valores de $\bar{X}$ y la línea central correspondiente a $\bar{\bar{X}}$. En otra gráfica mostrar los valores de R junto con la línea central correspondiente a $\bar{R}$.

Tabla 18.13

Subgrupo	Observaciones del subgrupo			
1	13	11	13	16
2	11	12	20	15
3	16	18	20	15
4	13	15	18	12
5	12	19	11	12
6	14	10	19	16
7	12	13	20	10
8	17	17	12	14
9	15	12	16	17
10	20	13	18	17

18.17 Una empresa de alimentos congelados elabora paquetes de ejotes de una libra (lb) (454 g). Cada hora se toman cuatro paquetes y se pesan con una exactitud de décimas de gramo. En la tabla 18.14 se presentan los datos obtenidos durante una semana.

Tabla 18.14

Lunes 10:00	Lunes 12:00	Lunes 2:00	Lunes 4:00	Martes 10:00	Martes 12:00	Martes 2:00	Martes 4:00	Miércoles 10:00	Miércoles 12:00
453.0	451.6	452.0	455.4	454.8	452.6	453.6	453.2	453.0	451.6
454.5	455.5	451.5	453.0	450.9	452.8	456.1	455.8	451.4	456.0
452.6	452.8	450.8	454.3	455.0	455.5	453.9	452.0	452.5	455.0
451.8	453.5	454.8	450.6	453.6	454.8	454.8	453.5	452.1	453.0

Miércoles 2:00	Miércoles 4:00	Jueves 10:00	Jueves 12:00	Jueves 2:00	Jueves 4:00	Viernes 10:00	Viernes 12:00	Viernes 2:00	Viernes 4:00
454.7	451.1	452.2	454.0	455.7	455.3	454.2	451.1	455.7	450.7
451.4	452.6	448.9	452.8	451.8	452.4	452.9	453.8	455.3	452.5
450.9	448.5	455.3	455.5	451.2	452.3	451.5	452.4	455.4	454.1
455.8	454.4	453.9	453.8	452.8	452.3	455.8	454.3	453.7	454.2

Usar el método visto en el problema 18.4 para estimar la desviación estándar conjuntando las varianzas de las 20 muestras. Usar esta estimación para hallar los límites de control de la gráfica X-barra. ¿Se encuentra alguna de las 20 medias de los subgrupos fuera de los límites de control?

18.18 Los límites de control en la gráfica R de los datos de la tabla 18.14 son LCL = 0 y UCL = 8.205. ¿Se encuentra alguno de los rangos de los subgrupos fuera de los límites 3 sigma?

18.19 El proceso del problema 18.17 mediante el cual se llenan paquetes de ejotes de 1 lb se modifica con objeto de reducir la variabilidad de los pesos de los paquetes. Después de haber empleado esta modificación durante algún tiempo, se vuelven a recolectar los datos de toda una semana y se grafican los rangos de los nuevos subgrupos usando los límites de control dados en el problema 18.18. En la tabla 18.15 se presentan los nuevos datos. ¿Parece haberse reducido a la variabilidad? Si se ha reducido la variabilidad, encontrar nuevos límites de control para la gráfica X empleando los datos de la tabla 18.15.

Tabla 18.15

Lunes 10:00	Lunes 12:00	Lunes 2:00	Lunes 4:00	Martes 10:00	Martes 12:00	Martes 2:00	Martes 4:00	Miércoles 10:00	Miércoles 12:00
454.9	454.2	454.4	454.7	454.3	454.2	454.6	453.6	454.4	454.6
452.7	453.6	453.6	453.9	454.2	452.8	454.5	453.2	455.0	454.1
457.0	454.4	453.6	454.6	454.2	453.3	454.3	453.6	454.6	453.3
454.2	453.9	454.3	453.9	453.4	453.3	454.9	453.1	454.1	454.3

Miércoles 2:00	Miércoles 4:00	Jueves 10:00	Jueves 12:00	Jueves 2:00	Jueves 4:00	Viernes 10:00	Viernes 12:00	Viernes 2:00	Viernes 4:00
453.0	453.9	453.8	455.1	454.2	454.4	455.1	455.7	452.2	455.4
454.0	454.2	453.3	453.3	453.0	452.6	454.6	452.8	453.7	452.8
452.9	454.3	454.1	454.7	453.8	454.9	454.1	453.8	454.4	454.7
454.2	454.7	454.7	453.9	453.9	454.2	454.6	454.9	454.5	455.1

PRUEBAS PARA CAUSAS ESPECIALES

18.20 Los operadores que hacen ajustes a las máquinas continuamente son un problema en los procesos industriales. La tabla 18.16 contiene un conjunto de datos (20 muestras cada una de tamaño 5) en las que éste es el caso. Encontrar los límites de control de la gráfica X-barra, elaborar la gráfica X-barra y hacer las ocho pruebas para causas especiales dadas en la tabla 18.3.

Tabla 18.16

1	2	3	4	5	6	7	8	9	10
2.006	2.001	1.993	1.983	2.003	1.977	1.972	1.998	2.015	1.985
1.994	1.982	1.989	1.983	2.024	1.966	1.988	1.992	2.000	1.983
1.981	1.996	2.012	1.991	2.005	1.996	2.001	2.005	2.026	1.994
2.000	1.972	2.025	1.970	1.996	1.985	2.026	1.985	2.006	2.010
1.997	2.006	2.027	2.001	2.009	1.991	2.014	1.966	2.012	2.031

11	12	13	14	15	16	17	18	19	20
2.010	1.982	2.002	1.993	2.024	1.980	2.008	1.982	2.017	1.992
1.986	1.985	2.011	1.978	2.015	2.004	1.975	2.025	1.982	1.997
2.004	1.997	2.002	1.982	2.029	2.006	2.024	1.972	2.004	2.010
2.000	1.991	2.014	2.005	2.016	2.007	1.990	1.981	2.029	1.990
2.012	1.979	2.013	1.999	1.999	1.982	1.996	1.984	2.004	2.003

CAPACIDAD DE PROCESOS

18.21 Supóngase que los límites de especificación para los paquetes de comida congelada del problema 18.17 son LSL = 450 g y USL = 458 g. Usar las estimaciones de μ y σ obtenidas en el problema 18.17 para hallar C_{PK}. Estimar también las ppm que no satisfacen las especificaciones.

18.22 En el problema 18.21, calcular el C_{PK} y estimar las ppm de no conformes después de las modificaciones hechas en el problema 18.19.

GRÁFICAS *P* Y *NP*

18.23 Una empresa produce fusibles para el sistema eléctrico de los automóviles. A lo largo de 30 días se prueban 500 fusibles por día. En la tabla 18.17 se presenta la cantidad de fusibles defectuosos hallados por día. Determinar la línea central y los límites superior e inferior de control de la gráfica *P*. ¿Parece estar el proceso bajo control estadístico? Si el proceso está bajo control estadístico, dar una estimación puntual de las tasas de ppm de defectuosos.

Tabla 18.17

Día	1	2	3	4	5	6	7	8	9	10
# de defectuosos	3	3	3	3	1	1	1	1	6	1
Día	11	12	13	14	15	16	17	18	19	20
# de defectuosos	1	1	5	4	6	3	6	2	7	3
Día	21	22	23	24	25	26	27	28	29	30
# de defectuosos	2	3	6	1	2	3	1	4	4	5

18.24 Supóngase que en el problema 18.23, el fabricante de fusibles decide usar una gráfica *NP* en lugar de una gráfica *P*. Encontrar la línea central y los límites superior e inferior de control de esta gráfica.

18.25 Scottie Long, el gerente del departamento de carnes de una cadena grande de supermercados, desea saber cuál es el porcentaje de paquetes de carne para hamburguesa que muestran ligera decoloración. Cada día se inspeccionan varios paquetes y se anota el número de ellos que muestra una pequeña decoloración. Estos datos se presentan en la tabla 18.18. Proporcionar los límites escalonados superiores de control de estos 20 subgrupos.

Tabla 18.18

Día	Tamaño del subgrupo	Cantidad de decolorados	Porcentaje de decolorados
1	100	1	1.00
2	150	1	0.67
3	100	0	0.00
4	200	1	0.50
5	200	1	0.50
6	150	0	0.00
7	100	0	0.00
8	100	0	0.00
9	150	0	0.00
10	200	2	1.00
11	100	1	1.00
12	200	1	0.50
13	150	3	2.00
14	200	2	1.00
15	150	1	0.67
16	200	1	0.50
17	150	4	2.67
18	150	0	0.00
19	150	0	0.00
20	150	2	1.33

OTRAS GRÁFICAS DE CONTROL

18.26 Antes de revisar este problema, leer el problema 18.11. Durante 24 horas se lee cada hora la temperatura de un horno que se usa para elaborar pan. La temperatura de horneado es crítica en el proceso y el horno trabaja sin interrupción a lo largo de todos los turnos. Estos datos se presentan en la tabla 18.19. Para vigilar la temperatura del proceso se emplea una gráfica de mediciones individuales. Encontrar la línea central y los rangos móviles correspondientes al uso de pares adyacentes de mediciones. ¿Cómo se encuentran los límites de control?

Tabla 18.19

Hora	1	2	3	4	5	6	7	8	9	10	11	12
Temperatura	350.0	350.0	349.8	350.4	349.6	350.0	349.7	349.8	349.4	349.8	350.7	350.9
Hora	13	14	15	16	17	18	19	20	21	22	23	24
Temperatura	349.8	350.3	348.8	351.6	350.0	349.7	349.8	348.6	350.5	350.3	349.1	350.0

18.27 Antes de revisar este problema, leer el problema 18.12. Usar MINITAB para elaborar una gráfica EWMA con los datos de la tabla 18.14. Usando como valor objetivo 454 g, ¿qué indica esta gráfica respecto al proceso?

18.28 Antes de revisar este problema, leer el problema 18.13 que se refiere a las gráficas de zonas. Con los datos de la tabla 18.16, elaborar una gráfica de zonas. ¿Indica la gráfica de zonas que haya algunas situaciones fuera de control? ¿Qué deficiencia de las gráficas de zonas muestra este problema?

18.29 Antes de revisar este problema, leer el problema 18.15. Dar los límites de control escalonados de la gráfica *U* del problema 18.15.

18.30 En el control de calidad se emplean también las *gráficas de Pareto*. Una gráfica de Pareto es una gráfica de barras en la que se enumeran los defectos observados en orden descendente. Los defectos que se encuentran con mayor frecuencia aparecen primero en la lista, seguidos por aquellos que se encuentran con menos frecuencia. Usando estas gráficas se pueden identificar áreas de problema para corregir aquellas causas a las que se debe el mayor porcentaje de defectos. En mascarillas para respiración inspeccionadas durante cierto tiempo, se encontraron los siguientes defectos: decoloración, tirante faltante, abolladuras, roturas y agujeros. En la tabla 18.20 se muestran los resultados.

Tabla 18.20

decoloración	decoloración	decoloración
tirante	tirante	tirante
decoloración	abolladura	tirante
decoloración	tirante	decoloración
tirante	decoloración	decoloración
decoloración	decoloración	abolladura
decoloración	abolladura	ruptura
ruptura	agujero	decoloración
abolladura	decoloración	agujero
decoloración	ruptura	ruptura

En la figura 18-13 se presenta una gráfica de Pareto generada con MINITAB. Los datos que aparecen en la tabla 18.20 se ingresan en la columna 1 de la hoja de cálculo. La secuencia para elaborar esta gráfica es la siguiente: "**Stat** → **Quality tools** → **Pareto charts**". De acuerdo con la gráfica de Pareto, ¿a qué tipo de defecto debe dársele mayor atención? ¿Cuáles son los dos tipos de defectos a los que debe dárseles mayor atención?

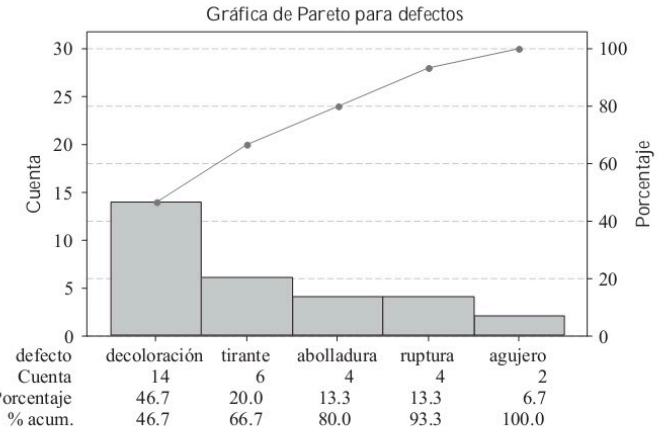


Figura 18-13 Defectos en mascarillas para respiración.

RESPUESTAS A LOS PROBLEMAS SUPLEMENTARIOS

CAPÍTULO 1

1.46 *a*) continua; *b*) continua; *c*) discreta; *d*) discreta; *e*) discreta.

1.47 *a*) Desde cero en adelante; continua. *b*) 2, 3, . . .; discreta.
c) Soltero, casado, divorciado, separado, viudo; discreta. *d*) Desde cero en adelante; continua.
e) 0, 1, 2, . . .; discreta.

1.48 *a*) 3 300; *b*) 5.8; *c*) 0.004; *d*) 46.74; *e*) 126.00; *f*) 4 000 000; *g*) 148; *h*) 0.000099; *i*) 2 180; *j*) 43.88.

1.49 *a*) 1 325 000; *b*) 0.0041872; *c*) 0.0000280; *d*) 7 300 000 000; *e*) 0.0003487; *f*) 18.50.

1.50 *a*) 3; *b*) 4; *c*) 7; *d*) 3; *e*) 8; *f*) una cantidad ilimitada; *g*) 3; *h*) 3; *i*) 4; *j*) 5.

1.51 *a*) 0.005 millones de bu o 5 000 bu; tres. *b*) 0.000000005 cm o 5×10^{-9} cm; cuatro. *c*) 0.5 ft; cuatro.
d) 0.05×10^{8} m o bien 5×10^{6} m; dos. *e*) 0.5 mi/s; seis. *f*) 0.5 millares de mi/s o bien 500 mi/s; tres.

1.52 *a*) 3.17×10^{-4}; *b*) 4.280×10^{8}; *c*) 2.160000×10^{4}; *d*) 9.810×10^{-6}; *e*) 7.32×10^{5}; *f*) 1.80×10^{-3}.

1.53 *a*) 374; *b*) 14.0.

1.54 *a*) 280 (dos cifras significativas), 2.8 centenas o 2.8×10^{2}; *b*) 178.9;
c) 250 000 (tres cifras significativas), 250 millares o 2.50×10^{5}; *d*) 53.0; *e*) 5.461; *f*) 9.05;
g) 11.54; *h*) 5 745 000 (cuatro cifras significativas), 5 745 millares, 5.745 millones o 5.745×10^{6}; *i*) 1.2;
j) 4 157.

1.55 *a*) -11; *b*) 2; *c*) $\frac{35}{8}$ o bien 4.375; *d*) 21; *e*) 3; *f*) -16; *g*) $\sqrt{98}$ o 9.89961 aproximadamente; *h*) $-7/\sqrt{34}$ o bien -1.20049 aproximadamente; *i*) 32; *j*) $10/\sqrt{17}$ o bien 2.42536 aproximadamente.

1.56 *a*) 22, 18, 14, 10, 6, 2, -2, -6 y -10; *b*) 19.6, 16.4, 13.2, 2.8, -0.8, -4 y -8.4; *c*) -1.2, 30, $10 - 4\sqrt{2} = 4.34$ aproximadamente y $10 + 4\pi = 22.57$ aproximadamente; *d*) 3, 1, 5, 2.1, -1.5, 2.5 y 0; *e*) $X = \frac{1}{4}(10 - Y)$.

1.57 *a*) -5; *b*) -24; *c*) 8.

1.58 *a*) -8; *b*) 4; *c*) -16.

1.76 *a*) -4; *b*) 2; *c*) 5; *d*) $\frac{3}{4}$; *e*) 1; *f*) -7.

1.77 *a*) $a = 3$, $b = 4$; *b*) $a = -2$, $b = 6$; *c*) $X = -0.2$, $Y = -1.2$; *d*) $A = \frac{184}{7} = 26.28571$ aproximadamente, $B = \frac{110}{7} = 15.71429$ aproximadamente; *e*) $a = 2$, $b = 3$, $c = 5$; *f*) $X = -1$, $Y = 3$, $Z = -2$; *g*) $U = 0.4$, $V = -0.8$, $W = 0.3$.

1.78 *b*) $(2, -3)$; es decir, $X = 2$, $Y = -3$.

1.79 *a*) 2, -2.5; *b*) 2.1 y -0.8 aproximadamente.

1.80 *a*) $\dfrac{4 \pm \sqrt{76}}{6}$ o bien 2.12 y -0.79 aproximadamente.
b) 2 y -2.5.
c) 0.549 y -2.549 aproximadamente.
d) $\dfrac{-8 \pm \sqrt{-36}}{2} = \dfrac{-8 \pm \sqrt{36}\sqrt{-1}}{2} = \dfrac{-8 \pm 6i}{2} = -4 \pm 3i$, donde $i = \sqrt{-1}$.
Estas raíces son *números complejos* y no se mostrarán cuando se emplee un procedimiento gráfico.

1.81 *a*) $-6.15 < -4.3 < -1.5 < 1.52 < 2.37$; *b*) $2.37 > 1.52 > -1.5 > -4.3 > -6.15$.

1.82 *a*) $30 \leq N \leq 50$; *b*) $S \geq 7$; *c*) $-4 \leq X < 3$; *d*) $P \leq 5$; *e*) $X - Y > 2$.

1.83 *a*) $X \geq 4$; *b*) $X > 3$; *c*) $N < 5$; *d*) $Y \leq 1$; *e*) $-8 \leq X \leq 7$; *f*) $-1.8 \leq N < 3$; *g*) $2 \leq a < 22$.

1.84 *a*) 1; *b*) 2; *c*) 3; *d*) -1; *e*) -2.

1.85 *a*) 1.0000; *b*) 2.3026; *c*) 4.6052; *d*) 6.9076; *e*) -2.3026.

1.86 *a*) 1; *b*) 2; *c*) 3; *d*) 4; *e*) 5.

1.87 Debajo de cada respuesta se muestra el comando de EXCEL.

```
1.160964   1.974636   2.9974102   1.068622    1.056642
=LOG(5,4)  =LOG(24,5) =LOG(215,6) =LOG(8,7)   =LOG(9,8)
```

1.88

```
> evalf(log[4](5));   1.160964047
> evalf(log[5](24));  1.974635869
> evalf(log[6](215)); 2.997410155
> evalf(log[7](8));   1.068621561
> evalf(log[8](9));   1.056641667
```

1.89 $\ln\left(\dfrac{a^3b^4}{c^5}\right) = 3\ln(a) + 4\ln(b) - 5\ln(c)$

1.90 $\log\left(\dfrac{xyz}{w^3}\right) = \log(x) + \log(y) + \log(z) - 3\log(w)$

1.91 $5\ln(a) - 4\ln(b) + \ln(c) + \ln(d) = \ln\left(\dfrac{a^5 cd}{b^4}\right)$.

1.92 $\log(u) + \log(v) + \log(w) - 2\log(x) - 3\log(y) - 4\log(z) = \log\left(\dfrac{uvw}{x^2 y^3 z^4}\right)$.

1.93 104/3.

1.94 2, −5/3.

1.95 $-\dfrac{5}{2} - \dfrac{\sqrt{7}}{2}i$ y $-\dfrac{5}{2} + \dfrac{\sqrt{7}}{2}i$.

1.96 165.13.

1.97 471.71.

1.98 402.14.

1.99 2.363.

1.100 0.617.

CAPÍTULO 2

2.19 *b*) 62.

2.20 *a*) 799; *b*) 1 000; *c*) 949.5; *d*) 1 099.5 y 1 199.5; *e*) 100 (horas); *f*) 76; *g*) $\frac{62}{400} = 0.155$ o 15.5%; *h*) 29.5%; *i*) 19.0%; *j*) 78.0%.

2.25 *a*) 24%; *b*) 11%; *c*) 46%.

2.26 *a*) 0.003 in; *b*) 0.3195, 0.3225, 0.3255, . . . , 0.3375 in.
c) 0.320-0.322, 0.323-0.325, 0.326-0.328, . . . , 0.335-0.337 in.

2.31 *a*) Cada una es de 5 años; *b*) cuatro (aunque estrictamente hablando el tamaño de la última clase no está especificado); *c*) uno; *d*) (85-94); *e*) 7 años y 17 años; *f*) 14.5 años y 19.5 años; *g*) 49.3% y 87.3%; *h*) 45.1%; *i*) no se puede determinar.

2.33 19.3, 19.3, 19.1, 18.6, 17.5, 19.1, 21.5, 22.5, 20.7, 18.3, 14.0, 11.4, 10.1, 18.6, 11.4 y 3.7. (Esto no suma 265 millones debido a los errores de redondeo en los porcentajes.)

2.34 *b*) 0.295, *c*) 0.19; *d*) 0.

CAPÍTULO 3

3.47 *a*) $X_1 + X_2 + X_3 + X_4 + 8$
b) $f_1 X_1^2 + f_2 X_2^2 + f_3 X_3^2 + f_4 X_4^2 + f_5 X_5^2$
c) $U_1(U_1 + 6) + U_2(U_2 + 6) + U_3(U_3 + 6)$
d) $Y_1^2 + Y_2^2 + \cdots + Y_N^2 - 4N$
e) $4X_1 Y_1 + 4Y_2 Y_2 + 4X_3 Y_3 + 4X_4 Y_4$.

3.48 *a*) $\sum_{j=1}^{3} (X_j + 3)^3$; *b*) $\sum_{j=1}^{15} f_j (Y_j - a)^2$; *c*) $\sum_{j=1}^{N} (2X_j - 3Y_j)$;

$$d) \sum_{j=1}^{8} \left(\frac{X_j}{Y_j} - 1 \right)^2; \qquad e) \frac{\sum_{j=1}^{12} f_j a_j^2}{\sum_{j=1}^{12} f_j}.$$

3.51 *a*) 20; *b*) −37; *c*) 53; *d*) 6; *e*) 226; *f*) −62; *g*) $\frac{25}{12}$.

3.52 *a*) −1; *b*) 23.

3.53 86.

3.54 0.50 s.

3.55 8.25.

3.56 *a*) 82; *b*) 79.

3.57 78.

3.58 66.7% varones y 33.3% mujeres.

3.59 11.09 tons.

3.60 501.0

3.61 0.72642 cm.

3.62 26.2.

3.63 715 min.

3.64 *b*) 1.7349 cm.

3.65 *a*) media = 5.4, mediana = 5; *b*) media = 19.91, mediana = 19.85.

3.66 85.

3.67 0.51 s.

3.68 8.

3.69 11.07 tons.

3.70 490.6.

3.71 0.72638 cm.

3.72 25.4.

3.73 Aproximadamente 78.3 años.

3.74 35.7 años.

3.75 708.3 min.

3.76 *a*) Media = 8.9, mediana = 9, moda = 7.
b) Media = 6.4, mediana = 6. Como cada uno de los números 4, 5, 6, 8 y 10 se presentan dos veces, se puede considerar que éstos son cinco modas; sin embargo, en este caso es más razonable concluir que no hay moda.

3.77 No existe una puntuación modal.

3.78 0.53 s.

3.79 10.

3.80 11.06 tons.

3.81 462.

3.82 0.72632 cm.

3.83 23.5.

3.84 668.7 min.

3.85 *a*) 35-39; *b*) 75 a 84.

3.86

a) Empleando la fórmula (9), moda = 11.1	Empleando la fórmula (10), moda = 11.3
b) Empleando la fórmula (9), moda = 0.7264	Empleando la fórmula (10), moda = 0.7263
c) Empleando la fórmula (9), moda = 23.5	Empleando la fórmula (10), moda = 23.8
d) Empleando la fórmula (9), moda = 668.7	Empleando la fórmula (10), moda = 694.9.

3.88 *a*) 8.4; *b*) 4.23.

3.89 *a*) $G = 8$; *b*) $\bar{X} = 12.4$.

3.90 *a*) 4.14; *b*) 45.8.

3.91 *a*) 11.07 tons; *b*) 499.5.

3.92 18.9%.

3.93 *a*) 1.01%; *b*) 238.2 millones; *c*) 276.9 millones.

3.94 $1 586.87.

3.95 $1 608.44.

3.96 3.6 y 14.4.

3.97 *a*) 3.0; *b*) 4.48.

3.98 *a*) 3; *b*) 0; *c*) 0.

3.100 *a*) 11.04; *b*) 498.2.

3.101 38.3 mi/h.

3.102 *b*) 420 mi/h.

3.104 *a*) 25; *b*) 3.55.

3.107 *a*) Cuartil inferior $= Q_1 = 67$, cuartil intermedio $= Q_2 =$ mediana $= 75$ y cuartil superior $= Q_3 = 83$.
b) 25% obtuvo 67 o menos (o 75% obtuvo 67 o más), 50% obtuvo 75 o menos (o 50% obtuvo 75 o más) y 75% obtuvo 83 o menos (o 25% obtuvo 83 o más).

3.108 *a*) $Q_1 = 10.55$ tons, $Q_2 = 11.07$ tons y $Q_3 = 11.57$ tons; *b*) $Q_1 = 469.3$, $Q_2 = 490.6$ y $Q_3 = 523.3$.

3.109 Media aritmética, mediana, moda, Q_2, P_{50} y D_5.

3.110 *a*) 10.15 tons; *b*) 11.78 tons; *c*) 10.55 tons; *d*) 11.57 tons.

3.112 *a*) 83; *b*) 64.

CAPÍTULO 4

4.33 *a*) 9; *b*) 4.273.

4.34 4.0 tons.

4.35 0.0036 cm.

4.36 7.88 kg.

4.37 20 semanas.

4.38 *a*) 18.2; *b*) 3.58; *c*) 6.21; *d*) 0; *e*) $\sqrt{2} = 1.414$ aproximadamente; *f*) 1.88.

4.39 *a*) 2; *b*) 0.85.

4.40 *a*) 2.2; *b*) 1.317.

4.41 0.576 ton.

4.42 *a*) 0.00437 cm; *b*) 60.0%, 85.2% y 96.4%.

4.43 *a*) 3.0; *b*) 2.8.

4.44 *a*) 31.2; *b*) 30.6.

4.45 *a*) 6.0; *b*) 6.0.

4.46 4.21 semanas.

4.48 *a*) 0.51 ton; *b*) 27.0; *c*) 12.

4.49 3.5 semanas.

4.52 *a*) 1.63 tons; *b*) 33.6 o 34.

4.53 El rango percentil 10-90 es igual a \$189 500 y el 80% de los precios de venta se encuentran en el intervalo \$130 250 $\pm$ \$94 750.

4.56 *a*) 2.16; *b*) 0.90; *c*) 0.484.

4.58 45.

4.59 *a*) 0.733 ton; *b*) 38.60; *c*) 12.1.

4.61 *a*) $\bar{X} = 2.47$; *b*) $s = 1.11$.

4.62 $s = 5.2$ y rango$/4 = 5$.

4.63 *a*) 0.00576 cm; *b*) 72.1%, 93.3% y 99.76%.

4.64 *a*) 0.719 ton; *b*) 38.24; *c*) 11.8.

4.65 *a*) 0.000569 cm; *b*) 71.6%, 93.0% y 99.68%.

4.66 *a*) 146.8 lb y 12.9 lb.

4.67 *a*) 1.7349 cm y 0.00495 cm.

4.74 *a*) 15; *b*) 12.

4.75 *a*) Estadística; *b*) álgebra.

4.76 *a*) 6.6%; *b*) 19.0%.

4.77 0.15.

4.78 0.20.

4.79 Álgebra.

4.80 0.19, −1.75, 1.17, 0.68, −0.29.

CAPÍTULO 5

5.15 *a*) 6; *b*) 40; *c*) 288; *d*) 2 188.

5.16 *a*) 0; *b*) 4; *c*) 0; *d*) 25.86.

5.17 *a*) −1; *b*) 5; *c*) −91; *d*) 53.

5.19 0, 26.25, 0, 1 193.1.

5.21 7.

5.22 *a*) 0, 6, 19, 42; *b*) −4, 22, −117, 560; *c*) 1, 7, 38, 155.

5.23 0, 0.2344, −0.0586, 0.0696.

5.25 *a*) $m_1 = 0$, *b*) $m_2 = pq$; *c*) $m_3 = pq(q - p)$; *d*) $m_4 = pq(p^2 - pq + q^2)$.

5.27 $m_1 = 0$, $m_2 = 5.97$, $m_3 = -0.397$, $m_4 = 89.22$.

5.29 m_1 (corregido) $= 0$, m_2 (corregido) $= 5.440$, m_3 (corregido) $= -0.5920$, m_4 (corregido) $= 76.2332$.

5.30 *a*) $m_1 = 0$, $m_2 = 0.53743$, $m_3 = 0.36206$, $m_4 = 0.84914$;
b) m_2 (corregido) $= 0.51660$, m_4 (corregido) $= 0.78378$.

5.31 *a*) 0; *b*) 52.95; *c*) 92.35; *d*) 7 158.20; *e*) 26.2; *f*) 7.28; *g*) 739.58; *h*) 22 247; *i*) 706 428; *j*) 24 545.

5.32 *a*) −0.2464; *b*) −0.2464.

5.33 0.9190.

5.34 La primera distribución.

5.35 *a*) 0.040; *b*) 0.074.

5.36 *a*) −0.02; *b*) −0.13.

5.37

	Distribución		
Coeficiente de asimetría (o sesgo) de Pearson	1	2	3
Primer coeficiente	0.770	1	−0.770
Segundo coeficiente	1.094	0	−1.094

5.38 *a*) 2.62; *b*) 2.58.

5.39 *a*) 2.94; *b*) 2.94.

5.40 *a*) La segunda; *b*) la primera.

5.41 *a*) La segunda; *b*) ninguna; *c*) la primera.

5.42 *a*) Mayor que 1 875; *b*) igual a 1 875; *c*) menor que 1 875.

5.43 *a*) 0.313.

CAPÍTULO 6

6.40 *a*) $\frac{5}{26}$; *b*) $\frac{5}{36}$; *c*) 0.98; *d*) $\frac{2}{9}$; *e*) $\frac{7}{8}$.

6.41 *a*) Probabilidad de obtener un rey en la primera extracción, pero no en la segunda extracción.
b) Probabilidad de obtener un rey, ya sea en la primera extracción, en la segunda extracción o en ambas.
c) No obtener rey en la primera extracción o no obtener rey en la segunda o en ninguna (es decir, no obtener rey ni en la primera extracción ni en la segunda extracción).
d) Probabilidad de obtener un rey en la tercera extracción dado que se obtuvo rey en la primea extracción pero no en la segunda extracción.
e) No obtener rey en la primera ni en la segunda ni en la tercera extracción.
f) Probabilidad de obtener rey en la primera y en la segunda extracción o no obtener rey en la segunda extracción pero sí en la tercera extracción.

6.42 *a*) $\frac{1}{3}$; *b*) $\frac{3}{5}$; *c*) $\frac{11}{15}$; *d*) $\frac{2}{5}$; *e*) $\frac{4}{5}$.

6.43 *a*) $\frac{4}{25}$; *b*) $\frac{4}{75}$; *c*) $\frac{16}{25}$; *d*) $\frac{64}{225}$; *e*) $\frac{11}{15}$; *f*) $\frac{1}{5}$; *g*) $\frac{104}{225}$; *h*) $\frac{221}{225}$; *i*) $\frac{6}{25}$; *j*) $\frac{52}{225}$.

6.44 *a*) $\frac{29}{185}$; *b*) $\frac{2}{37}$; *c*) $\frac{118}{185}$; *d*) $\frac{52}{185}$; *e*) $\frac{11}{15}$; *f*) $\frac{1}{5}$; *g*) $\frac{86}{185}$; *h*) $\frac{182}{185}$; *i*) $\frac{9}{37}$; *j*) $\frac{26}{111}$.

6.45 *a*) $\frac{5}{18}$; *b*) $\frac{11}{36}$; *c*) $\frac{1}{36}$.

6.46 *a*) $\frac{47}{52}$; *b*) $\frac{16}{221}$; *c*) $\frac{15}{34}$; *d*) $\frac{13}{17}$; *e*) $\frac{210}{221}$; *f*) $\frac{10}{13}$; *g*) $\frac{40}{51}$; *h*) $\frac{77}{442}$.

6.47 $\frac{5}{18}$.

6.48 *a*) 81:44; *b*) 21:4.

6.49 $\frac{19}{42}$.

6.50 *a*) $\frac{2}{5}$; *b*) $\frac{1}{5}$; *c*) $\frac{4}{15}$; *d*) $\frac{13}{15}$.

6.51 *a*) 37.5%; *b*) 93.75%; *c*) 6.25%; *d*) 68.75%.

6.52 *a*)

X	0	1	2	3	4
$p(X)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

6.53 *a*) $\frac{1}{48}$; *b*) $\frac{7}{24}$; *c*) $\frac{3}{4}$; *d*) $\frac{1}{6}$.

6.54 *a*)

X	0	1	2	3
$p(X)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{30}$

6.55 *a*)

X	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$p(X)$*	1	3	6	10	15	21	25	27	27	25	21	15	10	6	3	1

*Todos los valores de $p(x)$ tienen un divisor de 216.

b) 0.532407.

6.56 $9.

6.57 $4.80 por día.

6.58 *A* contribuye con $12.50; *B* contribuye con $7.50.

6.59 *a*) 7; *b*) 590; *c*) 541; *d*) 10 900.

6.60 *a*) 1.2; *b*) 0.56; *c*) $\sqrt{0.56} = 0.75$ aproximadamente.

6.63 10.5.

6.64 *a*) 12; *b*) 2 520; *c*) 720;
d) `=PERMUT(4,2), =PERMUT(7,5), =PERMUT(10,3)`.

6.65 $n = 5$.

6.66 60.

6.67 *a*) 5 040; *b*) 720; *c*) 240.

6.68 *a*) 8 400; *b*) 2 520.

6.69 *a*) 32 805; *b*) 11 664.

6.70 26.

6.71 *a*) 120; *b*) 72; *c*) 12.

6.72 *a*) 35; *b*) 70; *c*) 45.
d) `=COMBIN(7,3), =COMBIN(8,4), =COMBIN(10,8).`

6.73 $n = 6$.

6.74 210.

6.75 840.

6.76 *a*) 42 000; *b*) 7 000.

6.77 *a*) 120; *b*) 12 600.

6.78 *a*) 150; *b*) 45; *c*) 100.

6.79 *a*) 17; *b*) 163.

6.81 2.95×10^{25}.

6.83 *a*) $\frac{6}{5\,525}$; *b*) $\frac{22}{425}$; *c*) $\frac{169}{425}$; *d*) $\frac{73}{5\,525}$.

6.84 $\frac{171}{1\,296}$.

6.85 *a*) 0.59049; *b*) 0.32805; *c*) 0.08866.

6.86 *b*) $\frac{3}{4}$; *c*) $\frac{7}{8}$.

6.87 *a*) 8; *b*) 78; *c*) 86; *d*) 102; *e*) 20; *f*) 142.

6.90 $\frac{1}{3}$.

6.91 1/3 838 380 (es decir, las posibilidades en contra de ganar son 3 838 379 contra 1).

6.92 *a*) 658 007 a 1; *b*) 91 389 a 1; *c*) 9 879 a 1.

6.93 *a*) 649 739 a 1; *b*) 71 192 a 1; *c*) 4 164 a 1; *d*) 693 a 1.

6.94 $\frac{11}{36}$.

6.95 $\frac{1}{4}$.

6.96

X	3	4	5	6	7	8	9	10	11	12
$p(X)$*	1	3	6	10	12	12	10	5	3	1

*Todos los valores de $p(x)$ tienen un divisor de 64.

6.97 7.5

6.98 70%.

6.99 $(0.5)(0.01) + (0.3)(0.02) + (0.2)(0.03) = 0.017$.

6.100 $\dfrac{0.2(0.03)}{0.017} = 0.35$.

CAPÍTULO 7

7.35 *a*) 5 040; *b*) 210; *c*) 126; *d*) 165; *e*) 6.

7.36 *a*) $q^7 + 7q^6p + 21q^5p^2 + 35q^4p^3 + 35q^3p^4 + 21q^2p^5 + 7qp^6 + p^7$
b) $q^{10} + 10q^9p + 45q^8p^2 + 120q^7p^3 + 210q^6p^4 + 252q^5p^5 + 210q^4p^6 + 120q^3p^7 + 45p^2p^8 + 10qp^9 + p^{10}$

7.37 *a*) $\frac{1}{64}$; *b*) $\frac{3}{32}$; *c*) $\frac{15}{64}$; *d*) $\frac{5}{16}$; *e*) $\frac{15}{64}$; *f*) $\frac{3}{32}$; *g*) $\frac{1}{64}$;
h) **Función de densidad de probabilidad**

```
Binomial with n=6 and p=0.5
        X       P(X = x )
        0       0.015625
        1       0.093750
        2       0.234375
        3       0.312500
        4       0.234375
        5       0.093750
        6       0.015625
```

7.38 *a*) $\frac{57}{64}$; *b*) $\frac{21}{32}$;
c) `1-BINOMDIST(1,6,0.5,1) o bien 0.890625, =BINOMDIST(3,6,0.5,1) = 0.65625.`

7.39 *a*) $\frac{1}{4}$; *b*) $\frac{5}{16}$; *c*) $\frac{11}{16}$; *d*) $\frac{5}{8}$.

7.40 *a*) 250; *b*) 25; *c*) 500.

7.41 *a*) $\frac{17}{162}$; *b*) $\frac{1}{324}$.

7.42 $\frac{64}{243}$.

7.43 $\frac{193}{512}$.

7.44 *a*) $\frac{32}{243}$; *b*) $\frac{192}{243}$; *c*) $\frac{40}{243}$; *d*) $\frac{242}{243}$;
e)

```
a       0.131691        =BINOMDIST(5,5,0.66667,0)
b       0.790128        =1-BINOMDIST(2,5,0.66667,1)
c       0.164606        =BINOMDIST(2,5,0.66667,0)
d       0.995885        =1-BINOMDIST(0,5,0.66667,0).
```

7.45 *a*) 42; *b*) 3.550; *c*) −0.1127; *d*) 2.927.

7.47 *a*) $Npq(q - p)$; *b*) $Npq(1 - 6pq) + 3N^2p^2q^2$.

7.49 *a*) 1.5 y −1.6; *b*) 72 y 90.

7.50 *a*) 75.4; *b*) 9.

7.51 *a*) 0.8767; *b*) 0.0786; *c*) 0.2991;
d)

```
a       0.8767328      =NORMSDIST(2.4)-NORMSDIST(-1.2)
b       0.0786066      =NORMSDIST(1.87)-NORMSDIST(1.23)
c       0.2991508      =NORMSDIST(-0.5)-NORMSDIST(-2.35).
```

7.52 *a*) 0.0375; *b*) 0.7123; *c*) 0.9265; *d*) 0.0154; *e*) 0.7251; *f*) 0.0395;
g)

```
a       0.037538       =NORMSDIST(-1.78)
b       0.7122603      =NORMSDIST(0.56)
c       0.9264707      =1-NORMSDIST(-1.45)
d       0.0153863      =1-NORMSDIST(2.16)
e       0.7251362      =NORMSDIST(1.53)-NORMSDIST(-0.8)
f       0.0394927      =NORMSDIST(-2.52)+(1-NORMSDIST(1.83)).
```

7.53 *a*) 0.9495; *b*) 0.9500; *c*) 0.6826.

7.54 *a*) 0.75; *b*) -1.86; *c*) 2.08; *d*) 1.625 o bien 0.849; *e*) ± 1.645.

7.55 -0.995.

7.56 *a*) 0.0317; *b*) 0.3790; *c*) 0.1989;
d)

```
a       0.03174 =NORMDIST(2.25,0,1,0)
b       0.37903 =NORMDIST(-0.32,0,1,0)
c       0.19886 =NORMDIST(-1.18,0,1,0).
```

7.57 *a*) 4.78%; *b*) 25.25%; *c*) 58.89%.

7.58 *a*) 2.28%; *b*) 68.27%; *c*) 0.14%.

7.59 84.

7.60 *a*) 61.7%; *b*) 54.7%.

7.61 *a*) 95.4%; *b*) 23.0%; *c*) 93.3%.

7.62 *a*) 1.15; *b*) 0.77.

7.63 *a*) 0.9962; *b*) 0.0687; *c*) 0.0286; *d*) 0.0558.

7.64 *a*) 0.2511; *b*) 0.1342.

7.65 *a*) 0.0567; *b*) 0.9198; *c*) 0.6404; *d*) 0.0079.

7.66 0.0089.

7.67 *a*) 0.04979; *b*) 0.1494; *c*) 0.2241; *d*) 0.2241; *e*) 0.1680; *f*) 0.1008.

7.68 *a*) 0.0838; *b*) 0.5976; *c*) 0.4232.

7.69 *a*) 0.05610; *b*) 0.06131.

7.70 *a*) 0.00248; *b*) 0.04462; *c*) 0.1607; *d*) 0.1033; *e*) 0.6964; *f*) 0.0620.

7.71 *a*) 0.08208; *b*) 0.2052; *c*) 0.2565; *d*) 0.2138; *e*) 0.8911; *f*) 0.0142.

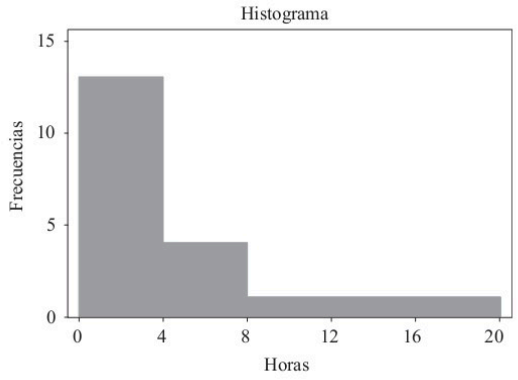
7.72 *a*) $\frac{5}{3\,888}$; *b*) $\frac{5}{324}$.

7.73 *a*) 0.0348; *b*) 0.000295.

7.74 $\frac{1}{16}$.

7.75 $p(X) = \binom{4}{X}(0.32)^X(0.68)^{4-X}$. Las frecuencias esperadas son 32, 60, 43, 13 y 2, respectivamente.

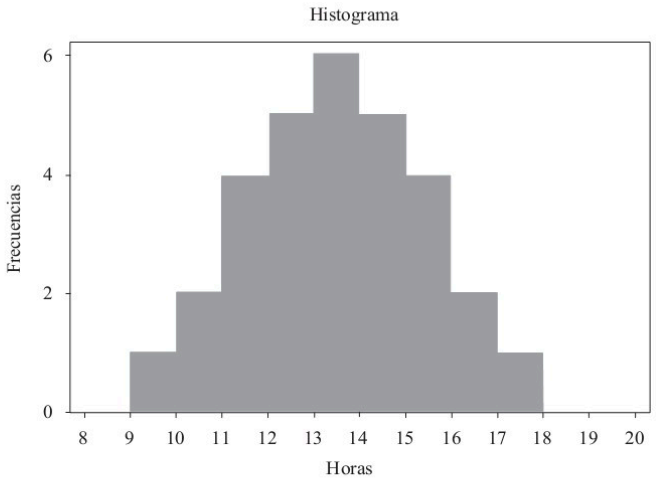
7.76



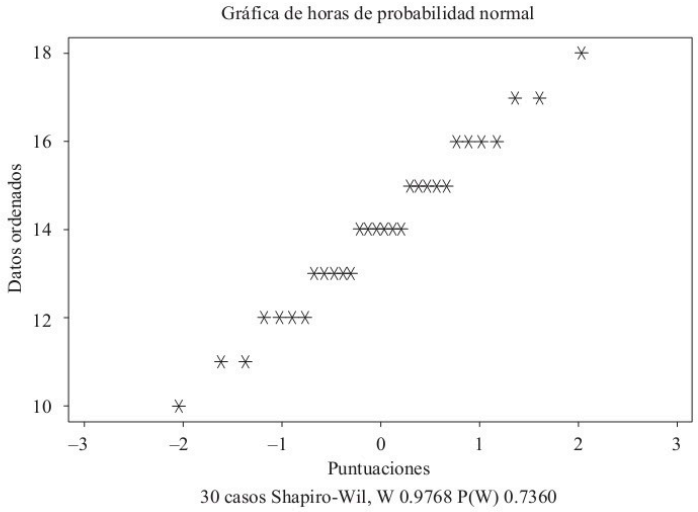
El histograma muestra un sesgo en los datos, lo que indica que no hay normalidad.

La prueba de Shapiro-Wilt de STATISTIX indica que no hay normalidad.

7.77 El histograma de STATISTIX indica claramente normalidad.

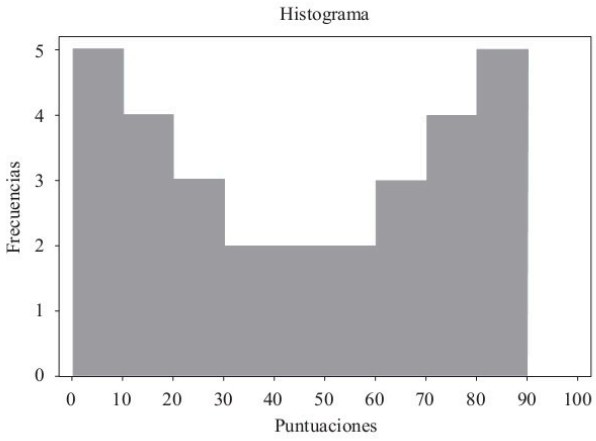


Con la secuencia "**Statistics** ⇒ **Randomness/Normality Tests** ⇒ **Normality Probability Plot**" se obtiene la gráfica siguiente. Si los datos provienen de una población distribuida normalmente, los puntos de la gráfica tienden a caer en una línea recta y $P(W)$ tiende a ser mayor que 0.05. Si $P(W) < 0.05$, por lo general se rechaza que haya normalidad.

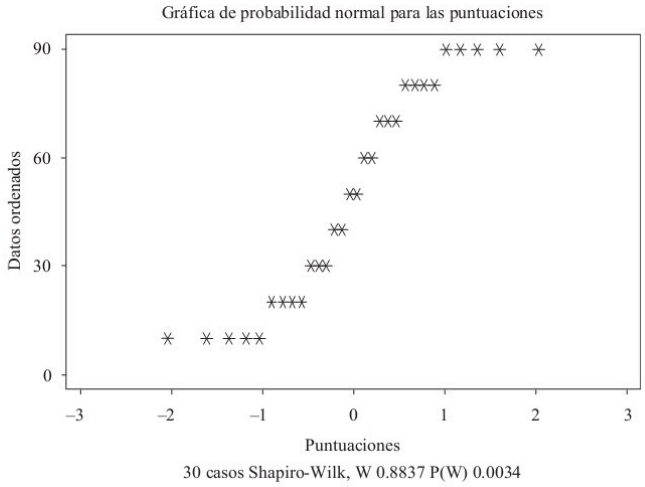


Arriba se muestra la gráfica para probabilidad normal. Se muestra el estadístico de Shapiro-Wilk junto con el valor p. $P(W) = 0.7360$. Como el valor p es considerablemente mayor que 0.05, no se rechaza la normalidad de los datos.

7.78 El siguiente histograma de las puntuaciones de examen de la tabla 7.11, obtenido con STATISTIX, tiene forma de U. Por lo tanto, se le conoce como distribución en forma de U.

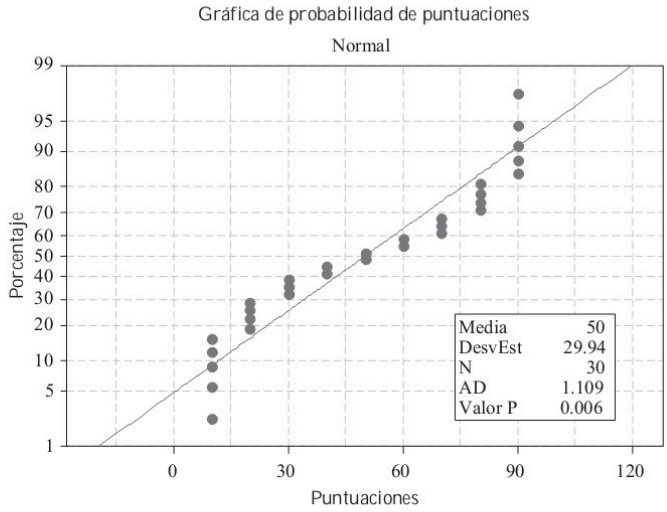


Esta gráfica se obtiene con la secuencia "**Statistics ⇒ Randomness/Normality Tests ⇒ Normality Probability Plot**". Si los datos provienen de una población distribuida normalmente, los puntos de la gráfica tienden a caer en una línea recta y $P(W)$ tiende a ser mayor que 0.05. Si $P(W) < 0.05$, por lo general se rechaza que haya normalidad.

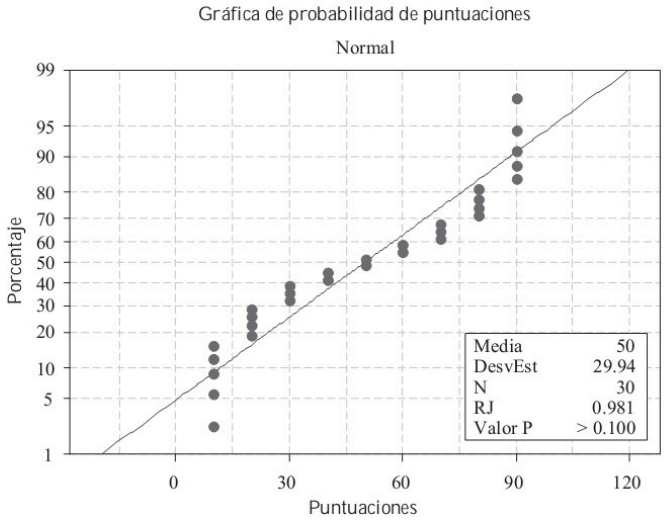


Arriba se muestra la gráfica para probabilidad normal. Se muestra el estadístico de Shapiro-Wilk junto con el valor p $P(W) = 0.0034$. Como el valor p es menor que 0.05, se rechaza la normalidad de los datos.

7.79 Además de la prueba de Kolgomorov-Smirnov de MINITAB y de la prueba de Shapiro-Wilk de STATISTIX, hay otras dos pruebas para la normalidad, que se verán aquí. Éstas son la prueba de Ryan-Joiner y la prueba de Anderson-Darling. Básicamente las cuatro pruebas calculan un estadístico de prueba y cada estadístico tiene un correspondiente valor p. Por lo general se sigue la regla siguiente. Si el valor p es < 0.05, se rechaza la normalidad. La gráfica siguiente se obtiene al hacer la prueba de Anderson-Darling. En este caso, el valor p es 0.006 y se rechazará la hipótesis de que los datos provienen de una distribución normal.



Obsérvese que si se emplea la prueba de Ryan-Joiner no se rechaza la normalidad.



7.80 $p(X) = \dfrac{(0.61)^X e^{-0.61}}{X!}$. Las frecuencias esperadas son 108.7, 66.3, 20.2, 4.1 y 0.7, respectivamente.

CAPÍTULO 8

8.21 *a*) 9.0; *b*) 4.47; *c*) 9.0; *d*) 3.16.

8.22 *a*) 9.0; *b*) 4.47; *c*) 9.0; *d*) 2.58.

8.23 *a*) $\mu_{\bar{X}} = 22.40$ g, $\sigma_{\bar{X}} = 0.008$ g; *b*) $\mu_{\bar{X}} = 22.40$ g, $\sigma_{\bar{X}} =$ un poco menos de 0.008 g.

8.24 *a*) $\mu_{\bar{X}} = 22.40$ g, $\sigma_{\bar{X}} = 0.008$ g; *b*) $\mu_{\bar{X}} = 22.40$ g, $\sigma_{\bar{X}} = 0.0057$ g.

8.25 *a*) 237; *b*) 2; *c*) ninguna; *d*) 34.

8.26 *a*) 0.4972; *b*) 0.1587; *c*) 0.0918; *d*) 0.9544.

8.27 *a*) 0.8164; *b*) 0.0228; *c*) 0.0038; *d*) 1.0000.

8.28 0.0026.

8.34 *a*) 0.0029; *b*) 0.9596; *c*) 0.1446.

8.35 *a*) 2; *b*) 996; *c*) 218.

8.36 *a*) 0.0179; *b*) 0.8664; *c*) 0.1841.

8.37 *a*) 6; *b*) 9; *c*) 2; *d*) 12.

8.39 *a*) 19; *b*) 125.

8.40 *a*) 0.0077; *b*) 0.8869.

8.41 *a*) 0.0028; *b*) 0.9172.

8.42 *a*) 0.2150; *b*) 0.0064; *c*) 0.4504.

8.43 0.0482.

8.44 0.0188.

8.45 0.0410.

8.47 *a*) 118.79 g; *b*) 0.74 g.

8.48 0.0228.

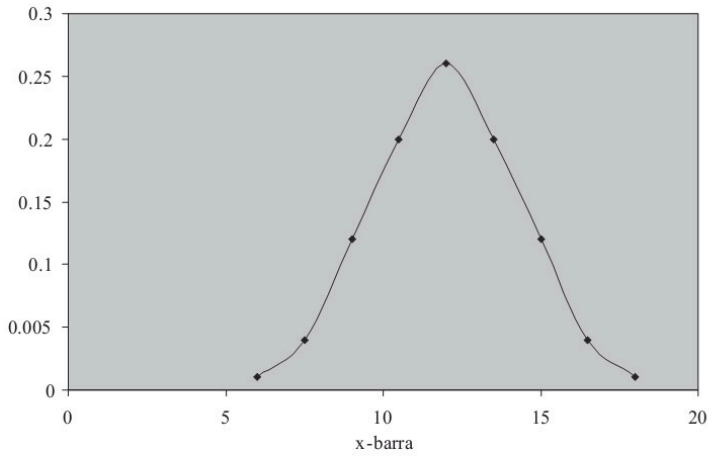
8.49 $\mu = 12$ y $\sigma^2 = 10.8$.

A	B	C	D
primera	segunda	media	probabilidad
6	6	6	0.01
6	9	7.5	0.02
6	12	9	0.04
6	15	10.5	0.02
6	18	12	0.01
9	6	7.5	0.02
9	9	9	0.04
9	12	10.5	0.08
9	15	12	0.04
9	18	13.5	0.02
12	6	9	0.04
12	9	10.5	0.08
12	12	12	0.16
12	15	13.5	0.08
12	18	15	0.04
15	6	10.5	0.02
15	9	12	0.04
15	12	13.5	0.08
15	15	15	0.04
15	18	16.5	0.02
18	6	12	0.01
18	9	13.5	0.02
18	12	15	0.04
18	15	16.5	0.02
18	18	18	0.01
			1

8.50 **Distribución de probabilidad de *X*-barra para $n = 2$.**

D	E	F	G	H
probabilidad		xbarra	p(xbarra)	
0.01		6	0.01	D2
0.02		7.5	0.04	D3+D7

0.04	9	0.12	D4+D8+D12
0.02	10.5	0.2	D5+D9+D13+D17
0.01	12	0.26	D6+D10+D14+D18+D22
0.02	13.5	0.2	D11+D15+D19+D23
0.04	15	0.12	D16+D20+D24
0.08	16.5	0.04	D21+D25
0.04	18	0.01	D26
0.02		1	SUM(G2:G10)
0.04			
0.08			
0.16			
0.08			
0.04			
0.02			
0.04			
0.08			
0.04			
0.02			
0.01			
0.02			
0.04			
0.02			
0.01			

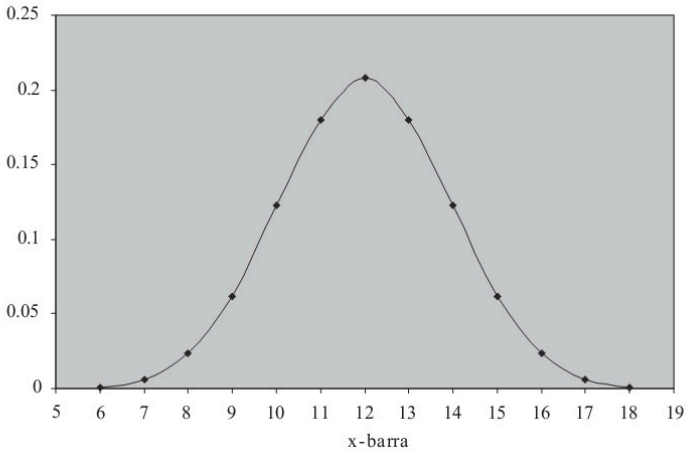


8.51 Media(x-barra) = 12 Var (x-barra) = 5.4.

8.52

xbarra	P(xbarra)
6	0.001
7	0.006
8	0.024
9	0.062

10	0.123
11	0.18
12	0.208
13	0.18
14	0.123
15	0.062
16	0.024
17	0.006
18	0.001



CAPÍTULO 9

9.21 *a*) 9.5 kg; *b*) 0.74 kg^2; *c*) 0.78 kg y 0.86 kg, respectivamente.

9.22 *a*) 1 200 h; *b*) 105.4 h.

9.23 *a*) Las estimaciones de las desviaciones estándar con muestras de tamaño 30, 50 y 100 cinescopios son 101.7 h, 101.0 h y 100.5 h, respectivamente; las estimaciones de las medias poblacionales son 1 200 h en todos los casos.

9.24 *a*) 11.09 ± 0.18 tons; *b*) 11.09 ± 0.24 tons.

9.25 *a*) 0.72642 ± 0.000095 in; *b*) 0.72642 ± 0.000085 in; *c*) 0.72642 ± 0.000072 in; *d*) 0.72642 ± 0.000060 in.

9.26 *a*) 0.72642 ± 0.000025 in; *b*) 0.000025 in.

9.27 *a*) Por lo menos 97; *b*) por lo menos 68; *c*) por lo menos 167; *d*) por lo menos 225.

9.28 Intervalo de confianza de 80% para la media: (286.064, 332.856).

9.29 *a*) $2\,400 \pm 45$ lb, $2\,400 \pm 59$ lb; *b*) 87.6%.

9.30 *a*) 0.70 ± 0.12, 0.69 ± 0.11; *b*) 0.70 ± 0.15, 0.68 ± 0.15; *c*) 0.70 ± 0.18, 0.67 ± 0.17.

9.31 Intervalo de confianza de 97.5% para la diferencia: (0.352477, 0.421523).

9.32 *a*) 16 400; *b*) 27 100; *c*) 38 420; *d*) 66 000.

9.33 *a*) 1.07 ± 0.09 h; *b*) 1.07 ± 0.12 h.

9.34 Intervalo de confianza de 85% para la diferencia: (−7.99550, −0.20948).

9.35 Intervalo de confianza de 95% para la diferencia: (−0.0918959, −0.00610414).

9.36 *a*) 180 ± 24.9 lb; *b*) 180 ± 32.8 lb; *c*) 180 ± 38.2 lb.

9.37

```
            One Sample Chi-square Test for a Variance
Sample Statistics for volume
        N         Mean       Std. Dev.   Variance
  ----------------------------------------------
       20        180.65      1.4677      2.1542
99% Confidence Interval for the variance
   Lower Limit    Upper Limit
   ---------     -----------
     1.06085      5.98045
```

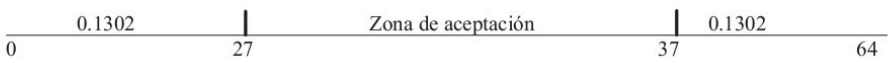
9.38

```
Two Sample Test for variances of units within line
Sample Statistics
    line
   Group       N        Mean        Std. Dev.     Variance
  ------------------------------------------------------
     1         13       104.9231    12.189        148.5769
     2         15       101.2667    5.5737        31.06667
  95% Confidence Interval of the Ratio of Two Variances
        Lower Limit      Upper Limit
          ---------      -----------
           1.568          15.334
```

CAPÍTULO 10

10.29 *a*) 0.2604.

b)



$\alpha = 0.1302 + 0.1302$.

10.30 *a*) Rechazar la hipótesis nula si $X \leq 21$ o $X \geq 43$, donde $X =$ número de canicas rojas extraídas; *b*) 0.99186; *c*) rechazar si $X \leq 23$ o $X \geq 41$.

10.31 *a*) H_o: $p = 0.5$ H_a: $p > 0.5$; *b*) prueba de una cola; *c*) rechazar la hipótesis nula si $X \geq 40$; *d*) rechazar la hipótesis nula si $X \geq 41$.

10.32 *a*) Para dos colas, valor $p =$ `2*(1-BINOMDIST(22,100,0.16666,1)` $= 0.126 > 0.05$. Al nivel de significancia 0.05, no se rechaza la hipótesis nula.
b) Para una cola, valor $p =$ `1-BINOMDIST(22,100,0.16666,1)` $= 0.063 > 0.05$. Al nivel de significancia 0.05, no se rechaza la hipótesis nula.

10.33 Empleando ya sea una prueba de una cola o una prueba de dos colas, al nivel de significancia 0.01 no se puede rechazar la hipótesis.

10.34 H_o: $p \geq 0.95$ H_a: $p < 0.95$
Valor $p = \mathrm{P}\{X \leq 182$ de 200 piezas cuando $p = 0.95\} =$ `BINOMDIST(182,200,0.95,1)` $= 0.012$. No se rechaza a 0.01, pero se rechaza a 0.05

10.35 Estadístico de prueba $= 2.63$, valores críticos ± 1.7805, se rechaza la hipótesis nula.

10.36 Estadístico de prueba $= -3.39$, valor crítico para 0.10 es -1.28155, valor crítico para 0.025 es -1.96. El resultado es significativo a $\alpha = 0.10$ y $\alpha = 0.025$.

10.37 Estadístico de prueba $= 6.46$, valor crítico $= 1.8808$, se concluye que $\mu > 25.5$.

10.38 `=NORMSINV(0.9)` $= 1.2815$ para $\alpha = 0.1$, `=NORMSINV(0.99)` $= 2.3263$ para $\alpha = 0.01$ y `=NORMSINV(0.999)` $= 3.0902$ para $\alpha = 0.001$.

10.39 valor $p = P\{X \leq 3\} + P\{X \geq 12\} = 0.0352$.

10.40 valor $p = P\{Z < -2.63\} + P\{Z > 2.63\} = 0.0085$.

10.41 valor $p = P\{Z < -3.39\} = 0.00035$.

10.42 valor $p = P\{Z > 6.46\} = 5.23515\text{E-}11$.

10.43 *a*) 8.64 ± 0.96 oz; *b*) 8.64 ± 0.83 oz; *c*) 8.64 ± 0.63 oz.

10.44 Los límites superiores de control son: *a*) 12 y *b*) 10.

10.45 Estadístico de prueba $= -5.59$, valor $p = 0.000$. Se rechaza la hipótesis nula ya que el valor $p < \alpha$.

10.46 Estadístico de prueba $= -1.58$, valor $p = 0.059$. Para $\alpha = 0.05$ no se rechaza, para $\alpha = 0.10$ sí se rechaza.

10.47 Estadístico de prueba $= -1.73$, valor $p = 0.042$. Para $\alpha = 0.05$ se rechaza, para $\alpha = 0.01$ no se rechaza.

10.48 Con una prueba de una cola se observa que, a ambos niveles de significancia, el nuevo fertilizante es mejor.

10.49 *a*) Estadístico de prueba $= 1.35$, valor $p = 0.176$, no se puede rechazar la hipótesis nula a $\alpha = 0.05$.
b) Estadístico de prueba $= 1.35$, valor $p = 0.088$, no se puede rechazar la hipótesis nula a $\alpha = 0.05$.

10.50 *a*) Estadístico de prueba $= 1.81$, valor $p = 0.07$, no se puede rechazar la hipótesis nula a $\alpha = 0.05$.
b) Estadístico de prueba $= 1.81$, valor $p = 0.0035$, se rechaza la hipótesis nula a $\alpha = 0.05$.

10.51 `=1-BINOMDIST(10,15,0.5,1)` o bien 0.059235.

10.52 =BINOMDIST(2,20,0.5,1) + 1−BINOMDIST(17,20,0.5,1) o 0.000402.

10.53 =BINOMDIST(10,15,0.6,1) o 0.7827.

10.54 =BINOMDIST(17,20,0.9,1)−BINOMDIST(2,20,0.9,1) o 0.3231.

10.55 El valor p se obtiene con =1−BINOMDIST(9,15,0.5,1) que da 0.1509. No se rechaza la hipótesis nula porque el valor de $\alpha = 0.0592$ y el valor p no es menor que α.

10.56 α=BINOMDIST(4,20,0.5,1) + 1−BINOMDIST(15,20,0.5,1) o bien 0.0118
valor p=BINOMDIST(3,20,0.5,1) + 1−BINOMDIST(16,20,0.5,1) o bien 0.0026.
Se rechaza la hipótesis nula dado que el valor $p < \alpha$.

10.57 =1−BINOMDIST(3,30,0.03,1) o bien 0.0119.

10.58 =BINOMDIST(3,30,0.04,1) o bien 0.9694.

10.59 α=1−BINOMDIST(5,20,0.16667,1) o bien 0.1018
valor p=1−BINOMDIST(6,20,0.16667,1) o bien 0.0371.

CAPÍTULO 11

11.20 *a*) 2.60; *b*) 1.75; *c*) 1.34; *d*) 2.95; *e*) 2.13.

11.21 *a*) 3.75; *b*) 2.68; *c*) 2.48; *d*) 2.39; *e*) 2.33.
a) =TINV(0.02,4) o bien 3.7469; *b*) 2.6810; *c*) 2.4851; *d*) 2.3901; *e*) 3.3515.

11.22 *a*) 1.71; *b*) 2.09; *c*) 4.03; *d*) −0.128.

11.23 *a*) 1.81; *b*) 2.76; *c*) −0.879; *d*) −1.37.

11.24 *a*) ±4.60; *b*) ±3.06; *c*) ±2.79; *d*) ±2.75; *e*) ±2.70.

11.25 *a*) 7.38 ± 0.79; *b*) 7.38 ± 1.11.
c) (6.59214,8.16786)(6.26825,8.49175).

11.26 *a*) 7.38 ± 0.70; *b*) 7.38 ± 0.92.

11.27 *a*) 0.289 ± 0.030 segundos; *b*) 0.298 ± 0.049 segundos.

11.28 Con una prueba de dos colas se observa que no hay evidencia, ni al nivel 0.05 ni al nivel 0.01, que indiquen que el tiempo medio de vida ha variado.

11.29 Con una prueba de una cola se observa que no hay disminución en la media al nivel 0.05 ni al nivel 0.01.

11.30 Con una prueba de dos colas se observa que el producto no satisface las especificaciones requeridas.

11.31 Con una prueba de una cola se observa, a ambos niveles, que el contenido de cobre es superior al requerido por las especificaciones.

11.32 Con una prueba de una cola se observa que la nueva máquina debe introducirse si el nivel de significancia que se adopte es 0.01, pero no debe introducirse si el nivel de significancia que se adopte es 0.05.

11.33 Con una prueba de una cola se observa que la marca *A* es mejor que la marca *B* al nivel de significancia 0.05.

11.34 Empleando una prueba de dos colas, al nivel de significancia 0.05, de acuerdo con las muestras no se puede concluir que haya diferencia entre la acidez de las dos soluciones.

11.35 Empleando una prueba de una cola, al nivel de significancia 0.05 se concluye que el primer grupo no es mejor que el segundo.

11.36 *a*) 21.0; *b*) 26.2; *c*) 23.3; *d*) `=CHIINV(0.05,12)` o bien 21.0261 `=CHIINV(0.01,12)` o bien 26.2170 `=CHIINV(0.025,12)` o bien 23.3367.

11.37 *a*) 15.5; *b*) 30.1; *c*) 41.3; *d*) 55.8.

11.38 *a*) 20.1; *b*) 36.2; *c*) 48.3; *d*) 63.7.

11.39 *a*) $\chi_1^2 = 9.59$ y $\chi_2^2 = 34.2$.

11.40 *a*) 16.0; *b*) 6.35; *c*) suponiendo áreas iguales en ambas colas, $\chi_1^2 = 2.17$ y $\chi_2^2 = 14.1$.

11.41 *a*) 87.0 a 230.9 h; *b*) 78.1 a 288.5 h.

11.42 *a*) 95.6 a 170.4 h; *b*) 88.9 a 190.8 h.

11.43 *a*) 122.5; *b*) 179.2; *c*) `=CHIINV(0.95,150)` o bien 122.6918; *d*) `=CHIINV(0.05,150)` o bien 179.5806.

11.44 *a*) 207.7; *b*) 295.2; *c*) `=CHIINV(0.975,250)` o bien 208.0978; *d*) `=CHIINV(0.025,250)` o bien 295.6886.

11.46 *a*) 106.1 a 140.5 h; *b*) 102.1 a 148.1 h.

11.47 105.5 a 139.6 h.

11.48 Con base en las muestras dadas, el aparente aumento de la variabilidad no es significativo a ninguno de los dos niveles.

11.49 La disminución aparente de la variabilidad es significativa al nivel 0.05, pero no al nivel 0.01.

11.50 *a*) 3.07; *b*) 4.02; *c*) 2.11; *d*) 2.83.

11.51 *a*) `=FINV(0.05,8,10)` o bien 3.0717; *b*) `=FINV(0.01,24,11)` o bien 4.0209; *c*) `=FINV(0.05,15,24)` o bien 2.1077; *d*) `=FINV(0.01,20,22)` o bien 2.8274.

11.52 Al nivel 0.05, la varianza de la muestra 1 es significativamente mayor, pero al nivel 0.01, no.

11.53 *a*) Sí; *b*) no.

CAPÍTULO 12

12.26 La hipótesis no puede rechazarse a ninguno de los dos niveles.

12.27 La conclusión es la misma que antes.

12.28 El nuevo profesor no sigue el patrón de notas de los otros profesores. (El que las notas sean mejores que el promedio *puede* deberse a una mejor capacidad para enseñar o a un estándar inferior, o ambas cosas.)

12.29 No hay razón para rechazar la hipótesis de que la moneda no esté cargada.

12.30 No hay razón para rechazar la hipótesis a ninguno de los dos niveles.

12.31 *a*) 10, 50 y 60, respectivamente;
b) al nivel de significancia 0.05 no se puede rechazar la hipótesis de que los resultados sean los esperados.

12.32 Al nivel de significancia 0.05, la diferencia es significativa.

12.33 *a*) El ajuste es bueno; *b*) no.

12.34 *a*) El ajuste es "muy bueno"; *b*) al nivel 0.05, el ajuste no es bueno.

12.35 *a*) Al nivel 0.05 el ajuste es muy malo; dado que la distribución binomial da un buen ajuste a los datos, esto coincide con el problema 12.33.
b) El ajuste es bueno, pero no "muy bueno".

12.36 Al nivel 0.05 se puede rechazar la hipótesis, pero no al nivel 0.01.

12.37 La conclusión es la misma que antes.

12.38 La hipótesis no se puede rechazar a ninguno de los dos niveles.

12.39 La hipótesis no se puede rechazar al nivel 0.05.

12.40 La hipótesis se puede rechazar a los dos niveles.

12.41 La hipótesis se puede rechazar a los dos niveles.

12.42 La hipótesis no se puede rechazar a ninguno de los dos niveles.

12.49 *a*) 0.3863 (no corregido) y 0.3779 (con la corrección de Yate).

12.50 *a*) 0.2205, 0.1985 (corregido); *b*) 0.0872, 0.0738 (corregido).

12.51 0.4651.

12.54 *a*) 0.4188, 0.4082 (corregido).

12.55 *a*) 0.2261, 0.2026 (corregido); *b*) 0.0875, 0.0740 (corregido).

12.56 0.3715.

CAPÍTULO 13

13.24 *a*) 4; *b*) 6; *c*) $\frac{28}{3}$; *d*) 10.5; *e*) 6; *f*) 9.

13.25 (2, 1).

13.26 *a*) $2X + Y = 4$; *b*) intersección con el eje $X = 2$, intersección con el eje $Y = 4$; *c*) $-2, -6$.

13.27 $Y = \frac{2}{3}X - 3$ o bien $2X - 3Y = 9$.

13.28 *a*) Pendiente $= \frac{3}{5}$, intersección con el eje $Y = -4$; *b*) $3X - 5Y = 11$.

13.29 *a*) $-\frac{4}{3}$; *b*) $\frac{32}{3}$; *c*) $4X + 3Y = 32$.

13.30 $X/3 + Y/(-5) = 1$ o bien $5X - 3Y = 15$.

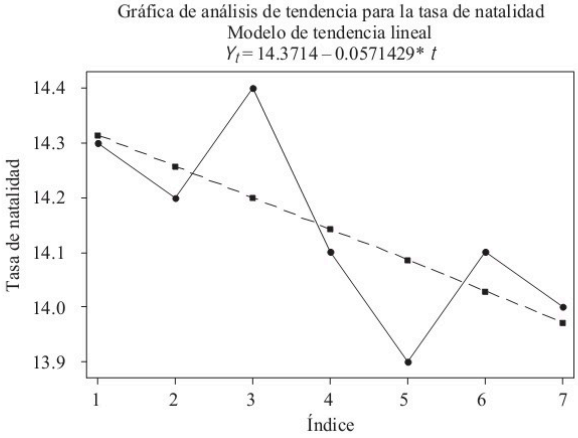
13.31 *a*) °F $= \frac{9}{5}$°C $+ 32$; *b*) 176°F; *c*) 20°C.

13.32 *a*) $Y = -\frac{1}{3} + \frac{5}{7}X$, o bien $Y = -0.333 + 0.714X$; *b*) $X = 1 + \frac{9}{7}Y$, o bien $X = 1.00 + 1.29Y$.

13.33 *a*) 3.24; 8.24; *b*) 10.00.

13.35 *b*) $Y = 29.13 + 0.661X$; *c*) $X = -14.39 + 1.15Y$; *d*) 79; *e*) 95.

13.36 *a*) y *b*).

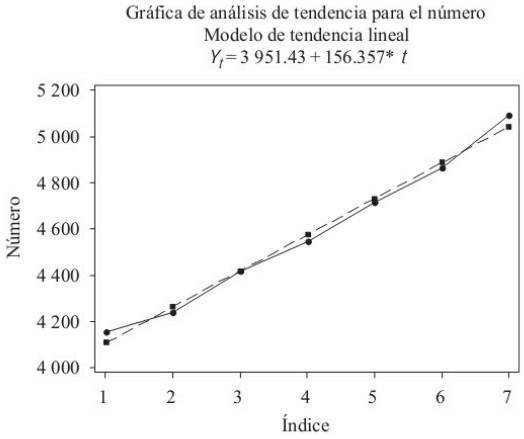


c)

Año	Tasa de natalidad	Valor ajustado	Residual
1998	14.3	14.3143	−0.014286
1999	14.2	14.2571	−0.057143
2000	14.4	14.2000	0.200000
2001	14.1	14.1429	−0.042857
2002	13.9	14.0857	−0.185714
2003	14.1	14.0286	0.071429
2004	14.0	13.9714	0.028571

d) $14.3714 - 0.0571429(13) = 13.6$.

13.37 *a*) y *b*)



c)

Año	Número	Valor ajustado	Residual
1999	4 154	4 107.79	46.2143
2000	4 240	4 264.14	−24.1429
2001	4 418	4 420.50	−2.5000
2002	4 547	4 576.86	−29.8571
2003	4 716	4 733.21	−17.2143
2004	4 867	4 889.57	−22.5714
2005	5 096	5 045.93	50.0714

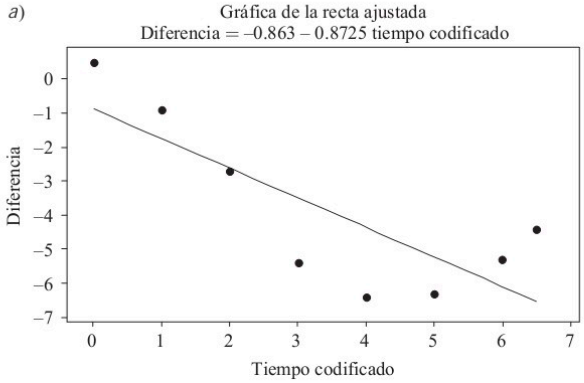
d) $3\,951.43 + 156.357(12) = 5\,827.7$.

13.38 $Y = 5.51 + 3.20(X - 3) + 0.733(X - 3)^2$ o bien $Y = 2.51 - 1.20X + 0.733X^2$.

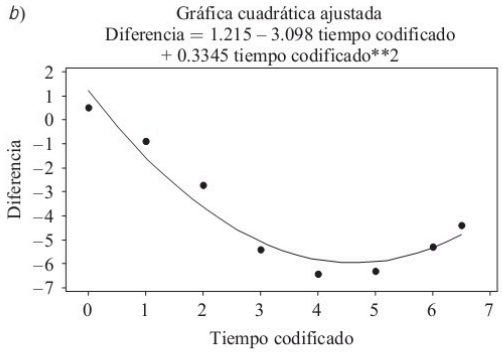
13.39 *b*) $D = 41.77 - 1.096V + 0.08786V^2$; *c*) 170 ft, 516 ft.

13.40

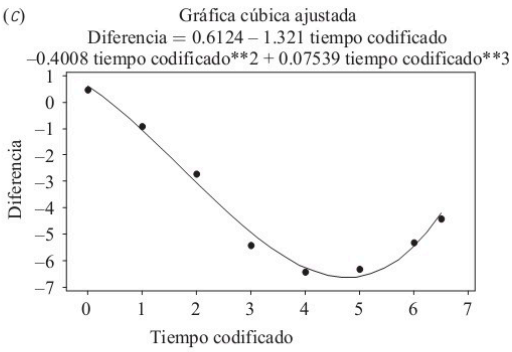
a)



b)



(*c*)



d)

	Modelo lineal		Modelo cuadrático		Modelo cúbico	
Año	Residual	Ajustado	Residual	Ajustado	Residual	Ajustado
1940	1.363	0.863	−0.715	1.215	−0.112	0.612
1950	0.836	1.736	0.649	−1.549	0.134	−1.034
1960	−0.092	0.092	0.943	−3.643	0.330	−3.030
1970	−1.919	1.919	−0.332	−5.068	−0.476	−4.924
1980	−2.047	2.047	−0.575	−5.825	−0.139	−6.261
1990	−1.074	1.074	−0.388	−5.912	0.292	−6.592
2000	0.798	6.098	0.030	−5.330	0.162	−5.462
2005	2.134	6.534	0.388	−4.788	−0.191	−4.209
	SSQ = 16.782		SSQ = 2.565		SSQ = 0.533	

e)

Lineal: $-0.863 - 0.8725(7) = -6.97$

Cuadrático: $1.215 - 3.098(7) + 0.3345(7^2) = -4.08$

Cúbico: $0.6124 - 1.321(7) - 0.4008(49) + 0.0754(343) = -2.41.$

13.41 *b*) Proporción = 0.965 + 0.0148 año codificado.

c)

Año	Año codificado	Hombre	Mujer	Proporción	Valor ajustado	Residual
1920	0	53.90	51.81	0.96	0.97	−0.00
1930	1	62.14	60.64	0.98	0.98	−0.00
1940	2	66.06	65.61	0.99	0.99	−0.00
1950	3	75.19	76.14	1.01	1.01	0.00
1960	4	88.33	90.99	1.03	1.02	0.01
1970	5	98.93	104.31	1.05	1.04	0.01
1980	6	110.05	116.49	1.06	1.05	0.00
1990	7	121.24	127.47	1.05	1.07	−0.02

d) Proporción pronosticada = 1.08. Proporción real = 1.04.

13.42 *b*) Diferencia = −2.63 + 1.35 x + 0.0064 xcuadrada.
d) La diferencia pronosticada para 1995 es −2.63 + 1.35(7.5) + 0.0064(56.25) = 7.86.

13.43 *b*) $Y = 32.14(1.427)^X$ o bien $Y = 32.14(10)^{0.1544X}$ o bien $Y = 32.14e^{0.3556X}$, donde $e = 2.718\cdots$ es la base logarítmica natural.
d) 387.

CAPÍTULO 14

14.40 *b*) $Y = 4.000 + 0.500X$; *c*) $X = 2.408 + 0.612Y$.

14.41 *a*) 1.304; *b*) 1.443.

14.42 *a*) 24.50; *b*) 17.00; *c*) 7.50.

14.43 0.5533.

14.44 Usando EXCEL la solución es `=CORREL(A2:A11,B2:B11)` que es 0.553.

14.45 1.5.

14.46 *a*) 0.8961; *b*) $Y = 80.78 + 1.138X$; *c*) 132.

14.47 *a*) 0.958; *b*) 0.872.

14.48 *a*) $Y = 0.8X + 12$; *b*) $X = 0.45Y + 1$.

14.49 *a*) 1.60; *b*) 1.20.

14.50 ± 0.80.

14.51 75%.

14.53 En los dos incisos se obtiene la misma respuesta, a saber -0.9203.

14.54 *a*) $Y = 18.04 - 1.34X$, $Y = 51.18 - 2.01X$.

14.58 0.5440.

14.59 *a*) $Y = 4.44X - 142.22$; *b*) 141.9 lb y 177.5 lb, respectivamente.

14.60 *a*) 16.92 lb; *b*) 2.07 in.

14.62 Correlación de Pearson entre `C1` y `C2` = 0.957.

14.63 Correlación de Pearson entre `C1` y `C2` = 0.582.

14.64 *a*) Sí; *b*) no.

14.65 *a*) No; *b*) sí.

14.66 *a*) 0.2923 y 0.7951; *b*) 0.1763 y 0.8361.

14.67 *a*) 0.3912 y 0.7500; *b*) 0.3146 y 0.7861.

14.68 *a*) 0.7096 y 0.9653; *b*) 0.4961 y 0.7235.

14.69 *a*) Sí; *b*) no.

14.70 *a*) 2.00 ± 0.21; *b*) 2.00 ± 0.28.

14.71 *a*) Usando una prueba de una cola, se puede rechazar la hipótesis.
b) Usando una prueba de una cola, no se puede rechazar la hipótesis.

14.72 *a*) 37.0 ± 3.28; *b*) 37.0 ± 4.45.

14.73 *a*) 37.0 ± 0.69; *b*) 37.0 ± 0.94.

14.74 *a*) 1.138 ± 0.398; *b*) 132.0 ± 16.6; *c*) 132.0 ± 5.4.

CAPÍTULO 15

15.26 *a*) $X_3 = b_{3.12} + b_{31.2}X_1 + b_{32.1}X_2$; *b*) $X_4 = b_{4.1235} + b_{41.235}X_1 + b_{42.135}X_2 + b_{43.125}X_3$.

15.28 *a*) $X_3 = 61.40 - 3.65X_1 + 2.54X_2$; *b*) 40.

15.29 *a*) $X_3 - 74 = 4.36(X_1 - 6.8) + 4.04(X_2 - 7.0)$ o bien $X_3 = 16.07 + 4.36X_1 + 4.04X_2$; *b*) 84 y 66.

15.30 En todos los casos se han resumido los resultados.

EXCEL

Price	Bedrooms	Baths
165	3	2
200	3	3
225	4	3
180	2	3
202	4	2
250	4	4
275	3	4
300	5	3
155	2	2
230	4	4

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.877519262
R Square	0.770040055
Adjusted R Square	0.704337213
Standard Error	25.62718211
Observations	10

	Coefficients
Intercept	**32.94827586**
Bedrooms	**28.64655172**
Baths	**29.28448276**

MINITAB

Análisis de regresión: precio contra recámaras, baños

The regression equation is

Price = 32.9 + 28.6 Bedrooms + 29.3 Baths

R-Sq = 77.0% R-Sq(adj) = 70.4%

SAS

Root MSE	25.62718	R-Square	0.7700
Dependent Mean	218.20000	Adj R-Sq	0.7043
Coeff Var	11.74481		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	32.94828	39.24247	0.84	0.4289
Bedrooms	Bedrooms	1	28.64655	9.21547	3.11	0.0171
Baths	Baths	1	29.28448	10.90389	2.69	0.0313

SPSS

Coeficientes[a]

Modelo		Coeficientes sin estandarizar		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	32.948	39.242		.840	.429
	Recámaras	28.647	9.215	.587	3.109	.017
	Baños	29.284	10.904	.507	2.686	.031

[a]Variable dependiente: Precio

STATISTIX

```
Statistix 8.0
Unweighted Least Squares Linear Regression of Price

Predictor
Variables     Coefficients     Std Error        T          P        VIF
Constant           32.9483       39.2425     0.84     0.4289
Bedrooms           28.6466       9.21547     3.11     0.0171        1.1
Baths              29.2845       10.9039     2.69     0.0313        1.1
R-Squared           0.7700

Estimated Price = 32.9 + 28.6(5) + 29.3(4) = 293.1 thousand.
```

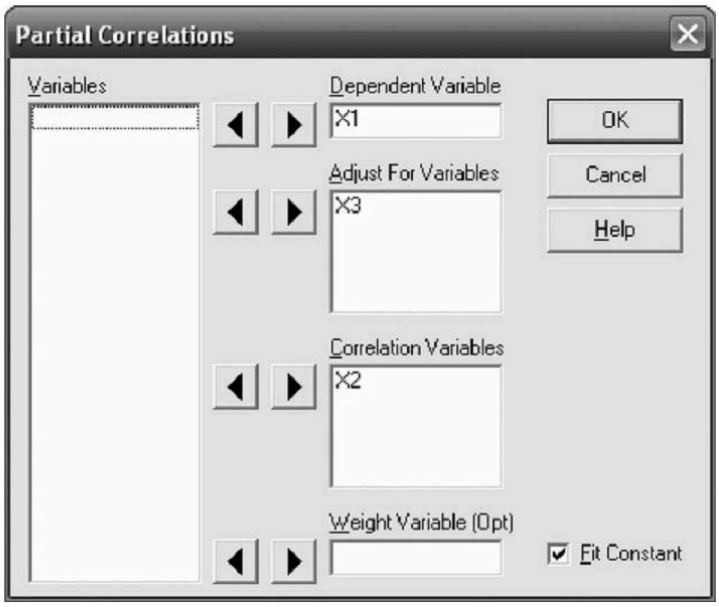
15.31 3.12.

15.32 *a*) 5.883; *b*) 0.6882.

15.33 0.9927.

15.34 *a*) 0.7567; *b*) 0.7255; *c*) 0.6710.

15.37 Se usa la secuencia "**Statistics** ⇒ **Linear models** ⇒ **Partial Correlations**". Se llena la siguiente ventana de diálogo como se muestra.



Se obtienen los resultados siguientes.

```
Statistix 8.0
Partial Correlations with X1
Controlled for X3
X2              0.5950
```

De igual manera, se encuentra que:

```
Statistix 8.0
Partial Correlations with X1
Controlled for X2

X3              -0.8995
```

y

```
Statistix 8.0
Partial Correlations with X2
Controlled for X1

X3               0.8727
```

15.38 *a*) 0.2672; *b*) 0.5099; *c*) 0.4026.

15.42 *a*) $X_4 = 6X_1 + 3X_2 - 4X_3 - 100$; *b*) 54.

15.43 *a*) 0.8710; *b*) 0.8587; *c*) −0.8426.

15.44 *a*) 0.8947; *b*) 2.680.

15.45 Con cualquiera de las soluciones siguientes se obtendrán los mismos coeficientes que resolviendo las ecuaciones normales. En EXCEL se usa la secuencia "**Tools** ⇒ **Data analysis** ⇒ **Regression**" para hallar la ecuación de regresión, así como otras medidas de regresión. La parte de los resultados a partir de la cual se obtiene la ecuación de regresión es la siguiente:

	Coeficientes
Intersección	–25.3355
Fumador	–302.904
Alcohol	–4.57069
Ejercicio	–60.8839
Alimentación	–36.8586
Peso	16.76998
Edad	–9.52833

En MINITAB se emplea la secuencia "**Stat** ⇒ **Regression** ⇒ **Regression**" para hallar la ecuación de regresión. La parte de los resultados a partir de la cual se encuentra la ecuación de regresión es la siguiente.

Análisis de regresión: Medcost *versus* fumador, alcohol, . . .

```
The regression equation is
Medcost = − 25 − 303 smoker − 4.6 alcohol − 60.9 Exercise − 37 Dietary + 16.8 weight
 − 9.53 Age

   Predictor      Coef     SE Coef          T       P
   Constant      −25.3       644.8      −0.04   0.970
   smoker       −302.9       256.1      −1.18   0.271
   alcohol       −4.57       11.89      −0.38   0.711
   Exercise     −60.88       19.75      −3.08   0.015
```

```
Dietary       −36.9       104.0       −0.35    0.732
weight       16.770       3.561        4.71    0.002
Age          −9.528       9.571       −1.00    0.349
```

En SAS se emplea la secuencia "**Statistics** ⇒ **Regression** ⇒ **Linear**" para hallar la ecuación de regresión. La parte de los resultados a partir de la cual se encuentra la ecuación de regresión es la siguiente.

```
                    The REG Procedure
                      Model: MODEL1
            Dependent Variable: Medcost Medcost

          Number of Observations Read                    16
          Number of Observations Used                    15
          Number of Observations with Missing Values      1

Root MSE            224.41971    R-Square    0.9029
Dependent Mean     2461.80000    Adj R-Sq    0.8301
Coeff Var             9.11608
```

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t value	Pr > |t|
Intercept	Intercept	1	−25.33552	644.84408	−0.04	0.9696
smoker	smoker	1	−302.90395	256.11003	−1.18	0.2709
alcohol	alcohol	1	−4.57069	11.88579	−0.38	0.7106
Exercise	Exercise	1	−60.88386	19.75371	−3.08	0.0151
Dietary	Dietary	1	−36.85858	104.04736	−0.35	0.7323
weight	weight	1	16.76998	3.56074	4.71	0.0015
Age	Age	1	−9.52833	9.57104	−1.00	0.3486

En SPSS se emplea la secuencia "**Analysis** ⇒ **Regression** ⇒ **Linear**" para hallar la ecuación de regresión. La parte de los resultados a partir de la cual se encuentra la ecuación de regresión es la siguiente. Ver bajo la columna de coeficientes no estandarizados.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Error estándar	Beta	t	Sig.
1	(Constante)	−25.336	644.844		−.039	.970
	fumador	−302.904	256.110	−.287	−1.183	.271
	alcohol	−4.571	11.886	−.080	−.385	.711
	ejercicio	−60.884	19.754	−.553	−3.082	.015
	dieta	−36.859	104.047	−.044	−.354	.732
	peso	−16.770	3.561	.927	4.710	.002
	edad	−9.528	9.571	−.124	−.996	.049

[a]Variable dependiente: medcost

En STATISTIX se emplea la secuencia "**Statistics** ⇒ **Linear models** ⇒ **Linear regression**" para hallar la ecuación de regresión. La parte de los resultados a partir de la cual se encuentra la ecuación de regresión es la siguiente. Ver bajo la columna de coeficientes.

Statistix 8.0

Unweighted Least Squares Linear Regression of Medcost

Predictor Variables	Coefficient	Std Error	T	P	VIF
Constant	**−25.3355**	644.844	−0.04	0.9696	
Age	**−9.52833**	9.57104	−1.00	0.3486	1.3
Dietary	**−36.8586**	104.047	−0.35	0.7323	1.3
Exercise	**−60.8839**	19.7537	−3.08	0.0151	2.6
alcohol	**−4.57069**	11.8858	−0.38	0.7106	3.6
smoker	**−302.904**	256.110	−1.18	0.2709	4.9
weight	**16.7700**	3.56074	4.71	0.0015	3.2

CAPÍTULO 16

16.21 A los niveles de significancia 0.05 y 0.01 no hay diferencia significativa entre las cinco variedades. El análisis proporcionado por MINITAB es el siguiente:

One-way ANOVA: A, B, C, D, E

Source	DF	SS	MS	F	P
Factor	4	27.2	6.8	0.65	0.638
Error	15	157.8	10.5		
Total	19	185.0			

S = 3.243 R-Sq = 14.71% R-Sq(adj) = 0.00%

Level	N	Mean	StDev	Individual 95% CIs For Mean Based on Pooled StDev
A	4	16.500	3.697	(-----------*-----------)
B	4	14.500	2.082	(-----------*-----------)
C	4	17.750	3.862	(-----------*-----------)
D	4	16.000	3.367	(-----------*-----------)
E	4	17.500	2.887	(-----------*-----------)

Pooled StDev = 3.243

```
---+---------+---------+---------+-----
12.0       15.0       18.0       21.0
```

16.22 A los niveles de significancia 0.05 y 0.01 no hay diferencia entre los cuatro tipos de neumáticos. El análisis proporcionado por STATISTIX es el siguiente:

Statistix 8.0

Completely Randomized AOV for Mileage

Source	DF	SS	MS	F	P
Type	3	77.500	25.8333	2.39	0.0992
Error	20	216.333	10.8167		
Total	23	293.833			

Grand Mean 34.083 CV 9.65

	Chi-Sq	DF	P
Bartlett's Test of Equal Variances	4.13	3	0.2476
Cochran's Q	0.5177		
Largest Var / Smallest Var	6.4000		

```
Component of variance for between groups      2.50278
Effective cell size                              6.0

Type        Mean
A         35.500
B         36.000
C         33.333
D         31.500
```

16.23 Al nivel de significancia 0.05 sí hay diferencia entre los tres métodos de enseñanza, pero no al nivel 0.01. El análisis proporcionado por EXCEL es el siguiente:

MétodoI	MétodoII	MétodoIII
75	81	73
62	85	79
71	68	60
58	92	75
73	90	81

Análisis de varianza de un factor

RESUMEN

Grupos	*Cuenta*	*Suma*	*Promedio*	*Varianza*
MétodoI	5	339	67.8	54.7
MétodoII	5	416	83.2	90.7
MétodoIII	5	368	73.6	67.8

Análisis de varianza

Origen de las variaciones	*SS*	*df*	*MS*	*F*	*Valor p*
Entre grupos	604.9333	2	302.4667	4.265098	0.040088
Dentro de los grupos	852.8	12	71.06667		
Total	1 457.733	14			

16.24 Al nivel de significancia 0.05 sí hay diferencia entre las cinco marcas, pero no al nivel 0.01. El análisis proporcionado por SPSS es el siguiente:

ANÁLISIS DE VARIANZA

mpg

	Suma de cuadrados	df	Cuadrado medio	F	Sig.
Entre grupos	52.621	4	13.155	4.718	.010
En los grupos	44.617	16	2.789		
Total	97.238	20			

16.25 A los dos niveles hay diferencia entre las cuatro materias. El análisis proporcionado por SAS es el siguiente:

```
                The ANOVA Procedure
                  Class Level Information
Class       Levels      Values

Subject          4      1 2 3 4

Number of Observations Read                  16
Number of Observations Used                  16
```

The ANOVA Procedure

Dependent Variable: Grade

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	365.5708333	121.8569444	7.35	0.0047
Error	12	198.8666667	16.5722222		
Corrected Total	15	564.4375000			

16.26 Al nivel de significancia 0.05 no hay diferencia ni entre los operadores ni entre las máquinas. A continuación se presenta el análisis proporcionado por MINITAB.

Two-way ANOVA: Number versus Machine, Operator

Source	DF	SS	MS	F	P
Machine	2	56	28.0	4.31	0.101
Operator	2	6	3.0	0.46	0.660
Error	4	26	6.5		
Total	8	88			

S = 2.550 R-Sq = 70.45% R-Sq(adj) = 40.91%

16.27 Al nivel de significancia 0.01 no hay diferencia ni entre los operadores ni entre las máquinas. A continuación se presenta el análisis proporcionado por EXCEL. Comparar el análisis de EXCEL con el proporcionado por MINITAB en el problema anterior.

	Operador1	Operador2	Operador3
Máquina1	23	27	24
Máquina2	34	30	28
Máquina3	28	25	27

Análisis de varianza de dos factores con una sola muestra por grupo

RESUMEN	*Cuenta*	*Suma*	*Promedio*	*Varianza*
Fila 1	3	74	24.66667	4.333333
Fila 2	3	92	30.66667	9.333333
Fila 3	3	80	26.66667	2.333333
Columna 1	3	85	28.33333	30.33333
Columna 2	3	82	27.33333	6.333333
Columna 3	3	79	26.33333	4.333333

Análisis de varianza

Origen de las variaciones	*SS*	*df*	*MS*	*F*	*Valor p*
Filas	56	2	28	4.307692	0.100535
Columnas	6	2	3	0.461538	0.660156
Error	26	4	6.5		
Total	88	8			

16.28 Al valor *p* en SPSS se le llama `sig`. El valor *p* para los bloques es 0.640 y el valor *p* para los tipos de maíz es 0.011, que es menor a 0.05 y por lo tanto es significativo. Al nivel de significancia 0.05 no hay diferencia entre los bloques. Al nivel de significancia 0.05 sí hay diferencias en los rendimientos debido al tipo de maíz. A continuación se presenta el análisis proporcionado por SPSS.

Pruebas de efectos entre temas

Variable dependiente: resultado

Origen	Tipo III Suma de cuadrados	df	Cuadrado medio	F	Sig.
Modelo corregido	77.600[a]	7	11.086	2.867	.053
Intersección	3 920.000	1	3 920.000	1 013.793	.000
bloque	10.000	4	2.500	.647	.640
tipo	67.600	3	22.533	5.828	.011
Error	46.400	12	3.867		
Total	4 044.000	20			
Total corregido	124.000	19			

[a]R cuadrada = .626 (R cuadrada ajustada = .408)

16.29 Al nivel de significancia 0.01, en los rendimientos no hay diferencias que se deban a los bloques o al tipo de maíz. Comparar estos resultados de STATISTIX con los resultados de SPSS del problema 16.28.

```
Statistix 8.0

Randomized Complete Block AOV Table for yield

Source     DF          SS          MS          F          P
block       4      10.000      2.5000       5.83     0.0108
type        3      67.600     22.5333
Error      12      46.400      3.8667
Total      19     124.000

Grand Mean 14.000           CV 14.05

Means of yield for type
type          Mean
1           13.600
2           17.000
3           12.000
4           13.400
```

16.30 SAS representa el valor p como `Pr > F`. En los dos últimos renglones de los resultados, se ve que tanto los automóviles como las marcas de los neumáticos son significativos al nivel 0.05.

```
                    The GLM Procedure
                Class Level Information
          Class          Levels        Values
          Auto                6        1 2 3 4 5 6
          Brand               4        1 2 3 4

                    The GLM Procedure
  Dependent Variable: lifetime
                                 Sum of
Source                 DF       Squares    Mean Square    F Value    Pr > F
Model                   8   201.3333333    25.1666667       4.08    0.0092
Error                  15    92.5000000     6.1666667
Corrected Total        23   293.8333333
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Auto	5	123.8333333	24.7666667	4.02	0.0164
Brand	3	77.5000000	25.8333333	4.19	0.0243

16.31 Compare el análisis de MINITAB en este problema con el de SAS en el problema 16.30. Al nivel de significancia 0.01, no hay diferencia entre los automóviles ni entre las marcas, ya que los valores p son 0.016 y 0.024, ambos mayores a 0.01.

Two-way ANOVA: lifetime versus Auto, Brand

Source	DF	SS	MS	F	P
Auto	5	123.833	24.7667	4.02	0.016
Brand	3	77.500	25.8333	4.19	0.024
Error	15	92.500	6.1667		
Total	23	293.833			

S = 2.483 R-Sq = 68.52% R-Sq(adj) = 51.73%

16.32 A continuación se presentan los resultados de STATISTIX. El valor p que es 0.3171 indica que, al nivel de significancia 0.05 no hay diferencia entre las escuelas.

Statistix 8.0

Randomized Complete Block AOV Table for Grade

Source	DF	SS	MS	F	P
Method	2	604.93	302.467		
School	4	351.07	87.767	1.40	0.3171
Error	8	501.73	62.717		
Total	14	1457.73			

Para los métodos de enseñanza, el valor de F es 4.82 y el valor p es 0.0423. Por lo tanto, al nivel de significancia 0.05, sí hay diferencia entre los métodos de enseñanza.

16.33 Los resultados de EXCEL indican que al nivel de significancia 0.05, ni el color del pelo ni las estaturas de las mujeres adultas tienen influencia en los logros escolares. El valor p para el color del pelo es 0.4534 y el valor p para la estatura es 0.2602.

	Pelirroja	Rubia	Castaña
Alta	75	78	80
Mediana	81	76	79
Baja	73	75	77

Análisis de varianza de dos factores con una sola muestra por grupo

RESUMEN	*Cuenta*	*Suma*	*Promedio*	*Varianza*
Fila 1	3	233	77.66667	6.333333
Fila 2	3	236	78.66667	6.333333
Fila 3	3	225	75	4
Columna 1	3	229	76.33333	17.33333
Columna 2	3	229	76.33333	2.333333
Columna 3	3	236	78.33333	2.333333

Análisis de varianza

Origen de las variaciones	*SS*	*df*	*MS*	*F*	*Valor p*
Filas	21.55556	2	10.77778	1.920792	0.260203
Columnas	10.88889	2	5.444444	0.970297	0.453378
Error	22.44444	4	5.611111		
Total	54.88889	8			

16.34 En los siguientes resultados de SPSS, al valor *p* se le llama `Sig`. El valor *p* para color de pelo es 0.453 y el valor *p* para estatura es 0.260. Éstos son los mismos valores *p* que se obtuvieron con EXCEL en el problema 16.33. Dado que ninguno de ellos es menor que 0.01, al nivel de significancia 0.01 no son significativos. Es decir, las puntuaciones no son diferentes de acuerdo con los distintos colores de pelo ni tampoco de acuerdo con las diferentes estaturas.

Tests of Between-Subjects Effects

Variable dependiente: puntuación

Origen	Tipo III Suma de cuadrados	df	Cuadrado medio	F	Sig.
Modelo corregido	32.444[a]	4	8.111	1.446	.365
Intersección	53 515.111	1	53 515.111	9 537.347	.000
Pelo	10.889	2	5.444	.970	.453
Estatura	21.556	2	10.778	1.921	.260
Error	22.444	4	5.611		
Total	53 570.000	9			
Total corregido	54.889	8			

[a]R cuadrada = .591 (R cuadrada ajustada = .182)

16.35 En los resultados de MINITAB se observa que al nivel de significancia 0.05 hay diferencias debidas a la ubicación, pero no hay diferencias debidas a los fertilizantes. La interacción es significativa al nivel 0.05.

ANOVA: yield versus location, fertilizer

```
Factor          Type     Levels     Values
location        fixed         2     1, 2
fertilizer      fixed         4     1, 2, 3, 4

Analysis of Variance for yield

Source                 DF        SS        MS         F       P
location                1    81.225    81.225     12.26   0.001
fertilizer              3    18.875     6.292      0.95   0.428
location*fertilizer     3    78.275    26.092      3.94   0.017
Error                  32   212.000     6.625
Total                  39   390.375
```

16.36 En los resultados de STATISTIX se observa que al nivel de significancia 0.01 hay diferencias debidas a la ubicación, pero no hay diferencias debidas a los fertilizantes. Al nivel 0.01 no hay una interacción significativa.

Statistix 8.0

Analysis of Variance Table for yield

Source	DF	SS	MS	F	P
fertilize	3	18.875	6.2917	0.95	0.4283
location	1	81.225	81.2250	12.26	0.0014
fertilize*location	3	78.275	26.0917	3.94	0.0169
Error	32	212.000	6.6250		
Total	39	390.375			

16.37 En los siguientes resultados de SAS, el valor p para las máquinas es 0.0664, el valor p para los operadores es 0.0004 y el valor p para la interacción es 0.8024. Al nivel de significancia 0.05, sólo los operadores son significativos.

The GLM Procedure
Class Level Information

Class	Levels	Values
Operator	4	1 2 3 4
Machine	2	1 2

Number of Observations Read	40
Number of Observations Used	40

The GLM Procedure
Dependent Variable: Articles

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	154.8000000	22.1142857	4.08	0.0027
Error	32	173.6000000	5.4250000		
Corrected Tot	39	328.4000000			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Machine	1	19.6000000	19.6000000	3.61	0.0664
Operator	3	129.8000000	43.2666667	7.98	0.0004
Operator*Machine	3	5.4000000	1.8000000	0.33	0.8024

Los siguientes resultados de MINITAB son iguales a los de SAS.

ANOVA: Articles versus Machine, Operator

Factor	Type	Levels	Values
Machine	fixed	2	1, 2
Operator	fixed	4	1, 2, 3, 4

Analysis of Variance for Articles

Source	DF	SS	MS	F	P
Machine	1	19.600	19.600	3.61	0.066
Operator	3	129.800	43.267	7.98	0.000
Machine*Operator	3	5.400	1.800	0.33	0.802
Error	32	173.600	5.425		
Total	39	328.400			

16.38 En los siguientes resultados de STATISTIX, los valores p para variaciones del suelo en dos direcciones perpendiculares son 0.5658 y 0.3633 y el valor p para los tratamientos es 0.6802. Al nivel de significancia 0.01, ninguno de los tres es significativo.

Statistix 8.0

Latin Square AOV Table for Yield

Source	DF	SS	MS	F	P
Row	3	17.500	5.8333	0.74	0.5658
Column	3	30.500	10.1667	1.28	0.3633
Treatment	3	12.500	4.1667	0.53	0.6802
Error	6	47.500	7.9167		
Total	15	108.000			

16.39 Los siguientes resultados de MINITAB son iguales a los resultados de STATISTIX del problema 16.38. Ninguno de los factores es significativo al nivel 0.05.

General Linear Model: Yield versus Row, Column, Treatment

Factor	Type	Levels	Values
Row	fixed	4	1, 2, 3, 4
Column	fixed	4	1, 2, 3, 4
Treatment	fixed	4	1, 2, 3, 4

Analysis of Variance for yield, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Row	3	17.500	17.500	5.833	0.74	0.567
Column	3	30.500	30.500	10.167	1.28	0.362
Treatment	3	12.500	12.500	4.167	0.53	0.680
Error	6	47.500	47.500	7.917		
Total	15	108.000				

16.40 En los resultados obtenidos con SPSS no se observa que al nivel de significancia 0.05 haya diferencia en los logros escolares debido al color de pelo, a la estatura o al lugar de nacimiento.

Pruebas de efectos entre temas

Variable dependiente: puntuación

Origen	Tipo III Suma de cuadrados	df	Cuadrado medio	F	Sig.
Modelo corregido	44.000[a]	6	7.333	1.347	.485
Intersección	53 515.111	1	53 515.11	9 829.306	.000
Pelo	10.889	2	5.444	1.000	.500
Estatura	21.556	2	10.778	1.980	.336
Por nacimiento	11.556	2	5.778	1.061	.485
Error	10.889	2	5.444		
Total	53 570.000	9			
Total corregido	54.889	8			

[a]R cuadrada = .802 (R cuadrada ajustada = .206)

16.41 En el análisis de MINITAB se observa que hay diferencias significativas debidas a las especies de los pollitos y a las cantidades de la primera sustancia química, pero no debidas a las cantidades de la segunda sustancia química o al peso inicial de los pollitos. Obsérvese que para las especies el valor p es 0.009 y para la primera sustancia química el valor p es 0.032.

General Linear Model: Wtgain versus Weight, Species, ...

Factor	Type	Levels	Values
Weight	fixed	4	1, 2, 3, 4
Species	fixed	4	1, 2, 3, 4
Chemical1	fixed	4	1, 2, 3, 4
Chemical2	fixed	4	1, 2, 3, 4

Analysis of Variance for Wtgain, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Weight	3	2.7500	2.7500	0.9167	2.20	0.267
Species	3	38.2500	38.2500	12.7500	30.60	0.009
Chemical1	3	16.2500	16.2500	5.4167	13.00	0.032
Chemical2	3	7.2500	7.2500	2.4167	5.80	0.091
Error	3	1.2500	1.2500	0.4167		
Total	15	65.7500				

16.42 En SPSS al valor p se le llama `Sig`. Hay diferencias significativas en la resistencia del cable debidas al tipo de cable, pero no hay diferencias significativas debidas a los operadores, las máquinas o las empresas.

Pruebas de efectos entre temas

Variable dependiente: resistencia

Origen	Tipo III Suma de cuadrados	df	Cuadrado medio	F	Sig.
Modelo corregido	6 579.750[a]	12	548.313	5.489	.094
Intersección	488 251.563	1	488 251.563	4 887.607	.000
Tipo	4 326.188	3	1 442.063	14.436	.027
Empresa	2 066.188	3	688.729	6.894	.074
Operador	120.688	3	40.229	.403	.763
Máquina	66.888	3	22.229	.223	.876
Error	299.688	3	99.896		
Total	495 131.000	16			
Total corregido	6 879.438	15			

[a]R cuadrada = .956 (R cuadrada ajustada = .782)

16.43 Al nivel de significancia 0.05 hay diferencias significativas entre los tres tratamientos, pero al nivel de significancia 0.01, no. A continuación se presenta el análisis que se obtiene con EXCEL.

A	B	C
3	4	6
5	2	4
4	3	5
4	3	5

Análisis de varianza de un factor

RESUMEN

Grupos	*Cuenta*	*Suma*	*Promedio*	*Varianza*
A	4	16	4	0.666667
B	4	12	3	0.666667
C	4	20	5	0.666667

Análisis de varianza

Origen de las variaciones	*SS*	*df*	*MS*	*F*	*Valor p*
Entre grupos	8	2	4	6	0.022085
Dentro de los grupos	6	9	0.666667		
Total	14	11			

16.44 El valor *p* que da MINITAB es 0.700. Entre los CI no hay diferencia debido a las estaturas.

One-way ANOVA: Tall, Short, Medium

```
Source   DF       SS      MS       F        P
Factor    2     55.8    27.9    0.37    0.700
Error    12    911.5    76.0
Total    14    967.3

S = 8.715   R-Sq = 5.77%   R-Sq(adj) = 0.00%

                                 Individual 95% CIs For Mean Based on
                                 Pooled StDev
Level     N      Mean    StDev  ----+---------+---------+---------+-----
Tall      5    107.00    10.58          (-----------*-----------)
Short     4    105.00     8.33       (-----------*-----------)
Medium    6    102.50     7.15    (-----------*-----------)
                                ----+---------+---------+---------+-----
                                  96.0      102.0     108.0     114.0
```

16.46 En los resultados de SPSS, al valor *p* se le llama `Sig`. Al nivel de significancia 0.05, existe una diferencia significativa en las puntuaciones de examen, debida tanto a ser o no veterano como al CI.

Pruebas de efectos entre temas

Variable dependiente: puntuación

Origen	Tipo III Suma de cuadrados	df	Cuadrado medio	F	Sig.
Modelo corregido	264.333[a]	3	88.111	176.222	.006
Intersección	38 080.667	1	38 080.667	76 161.333	.000
Veterano	24.000	1	24.000	48.000	.020
CI	240.333	2	120.167	240.333	.004
Error	1.000	2	.500		
Total	38 346.000	6			
Total corregido	265.333	5			

[a]R cuadrada = .996 (R cuadrada ajustada = .991)

16.47 En el análisis de STATISTIX se encuentra que al nivel de significancia 0.01 las diferencias en las puntuaciones de examen debidas a ser o no veterano no son significativas, pero las diferencias debidas al CI sí son significativas.

```
Statistix 8.0

Randomized Complete Block AOV Table for Score

Source       DF          SS          MS          F          P
Veteran       1      24.000      24.000       48.00     0.020
IQ            2     240.333     120.167      240.33    0.0041
```

```
Error         2      1.000      0.500
Total         5    265.333
```

16.48 En el análisis de MINITAB se observa que al nivel de significancia 0.05 las diferencias en las puntuaciones de examen debidas a la ubicación no son significativas, pero las diferencias debidas al CI, sí.

Two-way ANOVA: Testscore versus Location, IQ

```
Source     DF       SS       MS       F       P
Location    3    6.250    2.083    0.12   0.943
IQ          2  221.167  110.583    6.54   0.031
Error       6  101.500   16.917
Total      11  328.917
```

16.49 En el análisis de SAS se observa que al nivel de significancia 0.01 las diferencias en las puntuaciones de examen debidas a la ubicación no son significativas, pero las diferencias debidas al CI, sí. Recuerde que el valor *p* se escribe `Pr > F`.

```
                    The GLM Procedure
                Class Level Information
            Class       Levels      Values
            Location         4      1 2 3 4
            IQ               3      1 2 3

          Number of Observations Read                  12
          Number of Observations Used                  12

                    The GLM Procedure
Source        DF     Type III SS      Mean Square    F Value      Pr>F
Location       3       6.2500000        2.0833333       0.12    0.9430
IQ             2     221.1666667      110.5833333       6.54    0.0311
```

16.53 En el análisis de MINITAB se observa que debido a la ubicación, las cantidades de óxido no son significativas al nivel de significancia 0.05. No hay interacción significativa entre ubicación y sustancias químicas.

ANOVA: rust versus location, chemical

```
Factor       Type   Levels      Values
location     fixed       2      1, 2
chemical     fixed       3      1, 2, 3

Analysis of Variance for rust

Source               DF         SS         MS        F         P
location              1      0.667      0.667     0.67     0.425
chemical              2     20.333     10.167    10.17     0.001
location*chemical     2      0.333      0.167     0.17     0.848
Error                18     18.000      1.000
Total                23     39.333
```

16.54 En el análisis de STATISTIX se observa que al nivel de significancia 0.05, las diferencias en el rendimiento debidas a la ubicación son significativas, pero las diferencias debidas a las variedades no son significativas. No hay interacción significativa entre la ubicación y las variedades.

Statistix 8.0

Analysis of Variance Table for Yield

Source	*DF*	*SS*	*MS*	*F*	*P*
Variety	4	35.333	8.8333	1.07	0.3822
location	2	191.433	95.7167	11.60	0.0001
Variety*location	8	82.567	10.3208	1.25	0.2928
Error	45	371.250	8.2500		
Total	59	680.583			

16.55 En el análisis de MINITAB se observa que al nivel de significancia 0.01, las diferencias en el rendimiento debidas a la ubicación son significativas, pero las diferencias debidas a las variedades no son significativas. No hay interacción significativa entre la ubicación y las variedades.

General Linear Model: Yield versus location, Variety

Factor	Type	Levels	Values
location	fixed	3	1, 2, 3
variety	fixed	5	1, 2, 3, 4, 5

Analysis of Variance for Yield, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
location	2	191.433	191.433	95.717	11.60	0.000
Variety	4	35.333	35.333	8.833	1.07	0.382
location*Variety	8	82.567	82.567	10.321	1.25	0.293
Error	45	371.250	371.250	8.250		
Total	59	680.583				

16.56 Observando la ANOVA de SPSS y teniendo en cuenta que `Sig.` en SPSS es lo mismo que valor *p*, se ve que, al nivel 0.05, el `factor 1`, el `factor 2` y el `tratamiento` (treatment) no tienen un efecto significativo sobre la variable de respuesta, ya que el valor *p* de los tres es mayor que 0.05.

Pruebas de efectos entre temas

Variable dependiente: respuesta

Origen	Tipo III Suma de cuadrados	df	Cuadrado medio	F	Sig.
Modelo corregido	92.667[a]	6	15.444	7.316	.125
Intersección	2 567.111	1	2 567.111	1 216.000	.001
Factor1	74.889	2	37.444	17.737	.053
Factor2	17.556	2	8.778	4.158	.194
Tratamiento	.222	2	.111	.053	.950
Error	4.222	2	2.111		
Total	2 664.000	9			
Total corregido	96.889	8			

[a]R cuadrada = .956 (R cuadrada ajustada = .826)

16.58 Observando la ANOVA de SPSS y teniendo en cuenta que `Sig.` en SPSS es lo mismo que valor *p*, se ve que al nivel 0.05 el `Factor 1`, el `Factor 2`, el `tratamiento latino` (latin treatment) y el `tratamiento griego` (greek treatment) no tienen un efecto significativo sobre la variable de respuesta, ya que el valor *p* de los cuatro es mayor que 0.05.

Pruebas de efectos entre temas

Variable dependiente: respuesta

Origen	Tipo III Suma de cuadrados	df	Cuadrado medio	F	Sig.
Modelo corregido	362.750[a]	12	30.229	.924	.607
Intersección	1 914.063	1	1 914.063	58.482	.005
Factor1	5.188	3	1.729	.053	.981
Factor2	15.188	3	5.063	.155	.920
Latino	108.188	3	36.063	1.102	.469
Griego	234.188	3	78.063	2.385	.247
Error	98.188	3	32.729		
Total	2 375.000	16			
Total corregido	460.938	15			

[a]R cuadrada = .787 (R cuadrada ajustada = .065)

CAPÍTULO 17

17.26 Con la alternativa de dos colas, el valor *p* es 0.0352. Al nivel de significancia 0.05 hay diferencia debida al aditivo, pero al nivel 0.01, no. El valor *p* se obtiene usando la distribución binomial y no la aproximación normal a la binomial.

17.27 Con la alternativa de una cola, el valor *p* es 0.0176. Como el valor *p* es menor que 0.05, se rechaza la hipótesis nula de que no hay diferencia debida al aditivo.

17.28 Empleando EXCEL, el valor *p* se obtiene con `=1-BINOMDIST(24,31,0.5,1)` que da 0.00044. El programa es efectivo al nivel de significancia 0.05.

17.29 Empleando EXCEL, el valor *p* se obtiene con `=1-BINOMDIST(15,22,0.5,1)`, que da 0.0262. El programa es efectivo al nivel de significancia 0.05.

17.30 Empleando EXCEL, el valor *p* se obtiene con `=1-BINOMDIST(16,25,0.5,1)`, que da 0.0539. Al nivel de significancia 0.05 no se puede concluir que la marca *B* se prefiera a la marca *A*.

17.31

BS = 25	BS = 30	BS = 35	BS = 40
41	16	11	6	1
37	12	7	2	−3
25	0	−5	−10	−15
43	18	13	8	3
42	17	12	7	2
28	3	−2	−7	−12
32	7	2	−3	−8
36	11	6	1	−4
27	2	−3	−8	−13
33	8	3	−2	−7

35	10	5	0	−5
24	−1	−6	−11	−16
22	−3	−8	−13	−18
34	9	4	−1	−6
28	3	−2	−7	−12
38	13	8	3	−2
46	21	16	11	6
41	16	11	6	1
27	2	−3	−8	−13
31	6	1	−4	−9
23	−2	−7	−12	−17
30	5	0	−5	−10
37	12	7	2	−3
36	11	6	1	−4
24	−1	−6	−11	−16

0.00154388	2*BINOMDIST(4,24,0.5,1)	Reject null
0.307456255	2*BINOMDIST(9,24,0.5,1)	Do not reject null
0.541256189	2*BINOMDIST(10,24,0.5,1)	Do not reject null
0.004077315	2*BINOMDIST(5,25,0.5,1)	Reject null

17.34 La suma de los rangos de la muestra menor = 141.5 y la suma de los rangos de la muestra mayor = 158.5. El valor p para dos colas = 0.3488. Al nivel de significancia 0.05 no se rechaza la hipótesis nula de no diferencia, pues el valor $p >$ 0.05.

17.35 En el problema 17.34, al nivel 0.01 no se puede rechazar la hipótesis nula en la prueba de una cola.

17.36 La suma de los rangos de la muestra menor = 132.5 y la suma de los rangos de la muestra mayor = 77.5. Para dos colas el valor $p = 0.0044$. La hipótesis nula de no diferencia se rechaza tanto a nivel 0.01 como a nivel 0.05, ya que el valor $p <$ 0.05.

17.37 Al nivel de significancia 0.05, el agricultor del problema 17.36 puede concluir que el trigo II da mayor rendimiento que el trigo I.

17.38 Para la marca *A*, la suma de los rangos = 86.0 y para la marca *B*, la suma de los rangos = 124.0. Para dos colas el valor $p = 0.1620$. *a*) Al nivel de significancia 0.05, no se rechaza la hipótesis nula de no diferencia entre las dos marcas *versus* hay diferencia, ya que el valor $p > 0.05$. *b*) Al nivel de significancia 0.05 no se puede concluir que la marca *B* sea mejor que la marca *A*, ya que para una cola valor p (0.081) > 0.05.

17.39 Sí, se puede emplear tanto la prueba *U* como la prueba de los signos para determinar si hay diferencia entre las dos máquinas.

17.41 3.

17.42 6.

17.46 *a*) 246; *b*) 168; *c*) 0.

17.47 *a*) 236; *b*) 115; *c*) 100.

17.49 $H = 2.59$, $DF = 4$, $P = 0.629$. A los niveles de significancia 0.05 y 0.01, no hay diferencia entre los rendimientos de las cinco variedades, ya que el valor p es mayor que 0.01 y que 0.05.

17.50 $H = 8.42$, $DF = 3$, $P = 0.038$. Al nivel de significancia 0.05 sí hay diferencia entre las cuatro marcas de neumáticos, pero no al nivel de significancia 0.01, ya que $0.01 <$ valor $p < 0.05$.

17.51 $H = 6.54$, $DF = 2$, $P = 0.038$. Al nivel de significancia 0.05 sí hay diferencia entre los tres métodos de enseñanza, pero no al nivel de significancia 0.01, ya que $0.01 <$ valor $p < 0.05$.

17.52 $H = 9.22$, $DF = 3$, $P = 0.026$. Al nivel de significancia 0.05 hay diferencia significativa entre las cuatro materias, pero no al nivel de significancia 0.01, ya que $0.01 <$ valor $p < 0.05$.

17.53 *a*) $H = 7.88$, $DF = 8$, $P = 0.446$. A los niveles de significancia 0.01 y 0.05 no hay diferencia significativa entre las duraciones de los tres cinescopios, ya que el valor $p > 0.01$ y 0.05.
b) $H = 2.59$, $DF = 4$, $P = 0.629$. A los niveles de significancia 0.01 y 0.05 no hay diferencia significativa entre las cinco variedades de trigo, ya que el valor $p > 0.01$ y 0.05.
c) $H = 5.70$, $DF = 3$, $P = 0.127$. A los niveles de significancia 0.01 y 0.05 no hay diferencia significativa entre las cuatro marcas de neumáticos, ya que el valor $p > 0.01$ y > 0.05.

17.54 *a*) $H = 5.65$, $DF = 2$, $P = 0.059$. A los niveles de significancia 0.01 y 0.05 no hay diferencia entre los tres métodos de enseñanza, ya que el valor $p > 0.01$ y > 0.05.
b) $H = 10.25$, $DF = 4$, $P = 0.036$. Al nivel de significancia 0.05 hay diferencia entre las cinco marcas de gasolina, pero no al nivel 0.01, ya que $0.01 <$ valor $p < 0.05$.
c) $H = 9.22$, $DF = 3$, $P = 0.026$. Al nivel de significancia 0.05 hay diferencia entre las cuatro materias, pero no al nivel 0.01, ya que $0.01 <$ valor $p < 0.05$.

17.55 *a*) 8; *b*) 10.

17.56 *a*) El número de rachas es $V = 10$.
b) La prueba de aleatoriedad está basada en la distribución normal estándar. La media es

$$\mu_V = \frac{2N_1N_2}{N_1 + N_2} + 1 = \frac{2(11)(14)}{25} + 1 = 13.32$$

y la desviación estándar es

$$\sigma_V = \sqrt{\frac{2(11)(14)\{2(11)(14) - 11 - 14\}}{25^2(24)}} = 2.41$$

La Z encontrada es

$$Z = \frac{10 - 13.32}{2.41} = -1.38$$

Empleando EXCEL, el valor p es `=2*NORMSDIST(-1.38)` que es 0.1676. Dado que el valor p es grande, no se duda de la aleatoriedad.

17.57 *a*) Aun cuando el número de corridas es menor que lo esperado, el valor p no es menor que 0.05. No se rechaza la aleatoriedad de la secuencia (*10*).

Prueba de corridas: moneda

```
Runs test for coin
Runs above and below K = 0.4
The observed number of runs = 7
The expected number of runs = 10.6
8 observations above K, 12 below
* N is small, so the following approximation may be invalid.
P-value = 0.084
```

b) El número de rachas es mayor que lo esperado. Se rechaza la aleatoriedad de la secuencia (*11*).

Prueba de corridas: moneda

```
Runs test for coin
Runs above and below K = 0.5
```

```
The observed number of runs = 12
The expected number of runs = 7
6 observations above K, 6 below
* N is small, so the following approximation may be invalid.
P-value = 0.002
```

17.58 *a*)

Sequence			Runs
a	a	b	2
a	b	a	3
b	a	a	2

b)

Sampling distribution

V	*f*
2	2
3	1

c)

Probability distribution

V	Pr{*V*}
2	0.667
3	0.333

17.59 Media = 2.333 Varianza = 0.222

17.60 *a*)

Sequence				Runs
a	a	b	b	2
a	b	a	b	4
a	b	b	a	3
b	b	a	a	2
b	a	b	a	4
b	a	a	b	3

Sampling distribution

V	*f*
2	2
3	2
4	2

Probability distribution

V	Pr{*V*}
2	0.333
3	0.333
4	0.333

Mean *V* 3
Variance *V* 0.667

b)

Sequence				Runs
a	b	b	b	2
b	a	b	b	3

b	b	a	b	3
b	b	b	a	3

Sampling distribution

V	f
2	1
3	3

Probability distribution

V	Pr{V}
2	0.25
3	0.75

Mean V	2.75
Variance V	0.188

c)

Sequence					Runs
a	b	b	b	b	2
b	a	b	b	b	3
b	b	a	b	b	3
b	b	b	a	b	3
b	b	b	b	a	2

Sampling distribution

V	f
2	2
3	3

Probability distribution

V	Pr{V}
2	0.4
3	0.6

Mean V	2.6
Variance V	0.24

17.61 *a*)

Sequence						Runs
a	a	b	b	b	b	2
a	b	a	b	b	b	4
a	b	b	a	b	b	4
a	b	b	b	a	b	4
a	b	b	b	b	a	3
b	a	a	b	b	b	3
b	a	b	a	b	b	5
b	a	b	b	a	b	5
b	a	b	b	b	a	4
b	b	a	a	b	b	3
b	b	a	b	a	b	5
b	b	a	b	b	a	4
b	b	b	b	a	a	2
b	b	b	a	b	a	4
b	b	b	a	a	b	3

b)

Sampling distribution

V	f
2	2
3	4
4	6
5	3

c)

Probability distribution

V	Pr{V}
2	0.133
3	0.267
4	0.4
5	0.2

Mean V 3.667
Variance V 0.888

17.62 Supóngase que los renglones se leen uno por uno. Es decir, primero el renglón 1, luego el renglón 2, luego el renglón 3 y finalmente el renglón 4.

Prueba de corridas: Calificaciones

```
Runs test for Grade
Runs above and below K = 69
The observed number of runs = 26
The expected number of runs = 20.95
21 observations above K, 19 below
P-value = 0.105
```

Al nivel de significancia 0.05 puede suponerse que las calificaciones se registraron en forma aleatoria.

17.63 Supóngase que los datos se registraron renglón por renglón.

```
Runs test for price
Runs above and below K = 11.36
The observed number of runs = 10
The expected number of runs = 13.32
14 observations above K, 11 below
P-value = 0.168
```

Al nivel de significancia 0.05 puede suponerse que los precios son aleatorios.

17.64 En los dígitos después del punto decimal, considérese que 0 representa a un dígito par y 1 representa un dígito non.

```
Runs test for digit
Runs above and below K = 0.473684
The observed number of runs = 9
The expected number of runs = 10.4737
9 observations above K, 10 below
* N is small, so the following approximation may be invalid.
P-value = 0.485
```

Al nivel de significancia 0.05 puede suponerse que los dígitos son aleatorios.

17.65 Al nivel de significancia 0.05 puede suponerse que los dígitos son aleatorios.

17.66 Usando la aproximación normal, el valor calculado para Z es -1.62. Empleando EXCEL, el valor p calculado es `=2*NORMSDIST(-1.62)` que es 0.105.

17.67 Usando la aproximación normal, el valor calculado para Z es -1.38. Empleando EXCEL, el valor p calculado es `=2*NORMSDIST(-1.38)` que es 0.168.

17.68 Usando la aproximación normal, el valor calculado para Z es -0.70. Empleando EXCEL, el valor p calculado es `=2*NORMSDIST(-0.70)` que es 0.484.

17.70 Correlación de rangos de Spearman = 1.0 y coeficiente de correlación de Pearson = 0.998.

CAPÍTULO 18

18.16 Medias de los subgrupos: 13.25 14.50 17.25 14.50 13.50 14.75 13.75 15.00 15.00 17.00
Rangos de los subgrupos: 5 9 5 6 8 9 10 5 5 7
$\bar{\bar{X}} = 14.85,\ \bar{R} = 6.9.$

18.17 La estimación conjunta de σ es 1.741. LCL = 450.7, UCL = 455.9. Ninguna de las medias de los subgrupos está fuera de los límites de control.

18.18 No.

18.19 La gráfica indica que ha disminuido la variabilidad. Los nuevos límites de control son LCL = 452.9 y UCL = 455.2. Después de la modificación, también parece que el proceso se ha centrado más próximo al valor objetivo.

18.20 Los límites de control son LCL = 1.980 y UCL = 2.017. Los periodos 4, 5 y 6 no satisfacen la prueba 5. Los periodos 15 a 20 no satisfacen la prueba 4. Cada uno de estos puntos es el último de 14 puntos consecutivos alternantes hacia arriba y hacia abajo.

18.21 $C_{PK} = 0.63$. ppm de no conformes = 32 487.

18.22 $C_{PK} = 1.72$. ppm de no conformes = menos de 1.

18.23 Línea central = 0.006133, LCL = 0, UCL = 0.01661. El proceso está bajo control. ppm = 6 133.

18.24 Línea central = 3.067, LCL = 0, UCL = 8.304.

18.25 0.032 0.027 0.032 0.024 0.024 0.027 0.032 0.032 0.027 0.024 0.032 0.024
0.027 0.024 0.027 0.024 0.027 0.027 0.027 0.027

18.26 Línea central = $\bar{\bar{X}} = 349.9$.

Rangos móviles: 0.0 0.2 0.6 0.8 0.4 0.3 0.1 0.4 0.4 0.9 0.2 1.1 0.5 1.5 2.8 1.6 0.3 0.1 1.2 1.9 0.2
1.2 0.9

Media de los anteriores rangos móviles = $\bar{R}_M = 0.765$.

Límites de control de la gráfica de mediciones individuales $\bar{\bar{X}} \pm 3(\bar{R}_M/d_2)$. d_2 es una constante de la gráfica de mediciones individuales que se obtiene de tablas de diversas fuentes y que en este caso es igual a 1.128. LCL = 347.9 y UCL = 352.0.

18.27 La gráfica EWMA indica que las medias del proceso se encuentran constantemente abajo del valor objetivo. Las medias de los subgrupos 12 y 13 caen abajo de los límites de control inferior. Las medias de los grupos después del 13 están arriba de los límites inferiores de control; pero la media del proceso sigue estando abajo del valor objetivo.

18.28 La gráfica de zonas no indica que haya alguna situación fuera de control. Sin embargo, como se ve en el problema 18.20, hay 14 puntos consecutivos alternando hacia arriba y hacia abajo. Dada la manera en que funciona una gráfica de zonas, ésta no indica esta situación.

18.29 Los 20 límites de control inferiores son: 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.38 0.00 0.00 0.00 0.00 0.00 0.38 0.00 0.00 0.38 0.00 0.38 0.38.

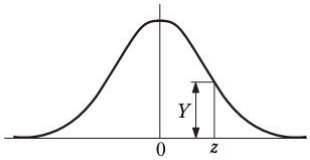
Los 20 límites de control superiores son: 9.52 9.52 9.52 9.52 9.52 7.82 8.46 7.07 7.82 8.46 9.52 9.52 9.52 7.07 9.52 9.52 7.07 9.52 7.07 7.07.

18.30 Decoloración; decoloración y pérdida del tirante.

Apéndices

Apéndice I

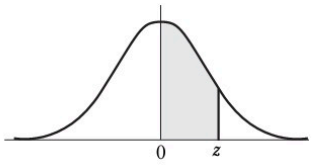
Ordenadas (*Y*) en *z*, en la curva normal estándar



z	0	1	2	3	4	5	6	7	8	9
0.0	.3989	.3989	.3989	.3988	.3986	.3984	.3982	.3980	.3977	.3973
0.1	.3970	.3965	.3961	.3956	.3951	.3945	.3939	.3932	.3925	.3918
0.2	.3910	.3902	.3894	.3885	.3876	.3867	.3857	.3847	.3836	.3825
0.3	.3814	.3802	.3790	.3778	.3765	.3752	.3739	.3725	.3712	.3697
0.4	.3683	.3668	.3653	.3637	.3621	.3605	.3589	.3572	.3555	.3538
0.5	.3521	.3503	.3485	.3467	.3448	.3429	.3410	.3391	.3372	.3352
0.6	.3332	.3312	.3292	.3271	.3251	.3230	.3209	.3187	.3166	.3144
0.7	.3123	.3101	.3079	.3056	.3034	.3011	.2989	.2966	.2943	.2920
0.8	.2897	.2874	.2850	.2827	.2803	.2780	.2756	.2732	.2709	.2685
0.9	.2661	.2637	.2613	.2589	.2565	.2541	.2516	.2492	.2468	.2444
1.0	.2420	.2396	.2371	.2347	.2323	.2299	.2275	.2251	.2227	.2203
1.1	.2179	.2155	.2131	.2107	.2083	.2059	.2036	.2012	.1989	.1965
1.2	.1942	.1919	.1895	.1872	.1849	.1826	.1804	.1781	.1758	.1736
1.3	.1714	.1691	.1669	.1647	.1626	.1604	.1582	.1561	.1539	.1518
1.4	.1497	.1476	.1456	.1435	.1415	.1394	.1374	.1354	.1334	.1315
1.5	.1295	.1276	.1257	.1238	.1219	.1200	.1182	.1163	.1145	.1127
1.6	.1109	.1092	.1074	.1057	.1040	.1023	.1006	.0989	.0973	.0957
1.7	.0940	.0925	.0909	.0893	.0878	.0863	.0848	.0833	.0818	.0804
1.8	.0790	.0775	.0761	.0748	.0734	.0721	.0707	.0694	.0681	.0669
1.9	.0656	.0644	.0632	.0620	.0608	.0596	.0584	.0573	.0562	.0551
2.0	.0540	.0529	.0519	.0508	.0498	.0488	.0478	.0468	.0459	.0449
2.1	.0440	.0431	.0422	.0413	.0404	.0396	.0387	.0379	.0371	.0363
2.2	.0355	.0347	.0339	.0332	.0325	.0317	.0310	.0303	.0297	.0290
2.3	.0283	.0277	.0270	.0264	.0258	.0252	.0246	.0241	.0235	.0229
2.4	.0224	.0219	.0213	.0208	.0203	.0198	.0194	.0189	.0184	.0180
2.5	.0175	.0171	.0167	.0163	.0158	.0154	.0151	.0147	.0143	.0139
2.6	.0136	.0132	.0129	.0126	.0122	.0119	.0116	.0113	.0110	.0107
2.7	.0104	.0101	.0099	.0096	.0093	.0091	.0088	.0086	.0084	.0081
2.8	.0079	.0077	.0075	.0073	.0071	.0069	.0067	.0065	.0063	.0061
2.9	.0060	.0058	.0056	.0055	.0053	.0051	.0050	.0048	.0047	.0046
3.0	.0044	.0043	.0042	.0040	.0039	.0038	.0037	.0036	.0035	.0034
3.1	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026	.0025	.0025
3.2	.0024	.0023	.0022	.0022	.0021	.0020	.0020	.0019	.0018	.0018
3.3	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014	.0013	.0013
3.4	.0012	.0012	.0012	.0011	.0011	.0010	.0010	.0010	.0009	.0009
3.5	.0009	.0008	.0008	.0008	.0008	.0007	.0007	.0007	.0007	.0006
3.6	.0006	.0006	.0006	.0005	.0005	.0005	.0005	.0005	.0005	.0004
3.7	.0004	.0004	.0004	.0004	.0004	.0004	.0003	.0003	.0003	.0003
3.8	.0003	.0003	.0003	.0003	.0003	.0002	.0002	.0002	.0002	.0002
3.9	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001

Apéndice II

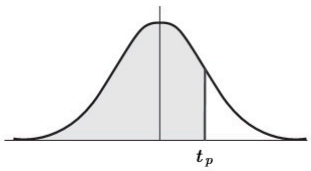
Áreas bajo la curva normal estándar, desde 0 hasta *z*



z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

Apéndice III

Valores percentiles (t_p) correspondientes a la distribución *t* de Student con ν grados de libertad (área sombreada = *p*)

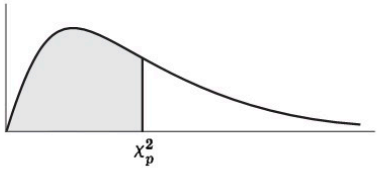


ν	$t_{.995}$	$t_{.99}$	$t_{.975}$	$t_{.95}$	$t_{.90}$	$t_{.80}$	$t_{.75}$	$t_{.70}$	$t_{.60}$	$t_{.55}$
1	63.66	31.82	12.71	6.31	3.08	1.376	1.000	.727	.325	.158
2	9.92	6.96	4.30	2.92	1.89	1.061	.816	.617	.289	.142
3	5.84	4.54	3.18	2.35	1.64	.978	.765	.584	.277	.137
4	4.60	3.75	2.78	2.13	1.53	.941	.741	.569	.271	.134
5	4.03	3.36	2.57	2.02	1.48	.920	.727	.559	.267	.132
6	3.71	3.14	2.45	1.94	1.44	.906	.718	.553	.265	.131
7	3.50	3.00	2.36	1.90	1.42	.896	.711	.549	.263	.130
8	3.36	2.90	2.31	1.86	1.40	.889	.706	.546	.262	.130
9	3.25	2.82	2.26	1.83	1.38	.883	.703	.543	.261	.129
10	3.17	2.76	2.23	1.81	1.37	.879	.700	.542	.260	.129
11	3.11	2.72	2.20	1.80	1.36	.876	.697	.540	.260	.129
12	3.06	2.68	2.18	1.78	1.36	.873	.695	.539	.259	.128
13	3.01	2.65	2.16	1.77	1.35	.870	.694	.538	.259	.128
14	2.98	2.62	2.14	1.76	1.34	.868	.692	.537	.258	.128
15	2.95	2.60	2.13	1.75	1.34	.866	.691	.536	.258	.128
16	2.92	2.58	2.12	1.75	1.34	.865	.690	.535	.258	.128
17	2.90	2.57	2.11	1.74	1.33	.863	.689	.534	.257	.128
18	2.88	2.55	2.10	1.73	1.33	.862	.688	.534	.257	.127
19	2.86	2.54	2.09	1.73	1.33	.861	.688	.533	.257	.127
20	2.84	2.53	2.09	1.72	1.32	.860	.687	.533	.257	.127
21	2.83	2.52	2.08	1.72	1.32	.859	.686	.532	.257	.127
22	2.82	2.51	2.07	1.72	1.32	.858	.686	.532	.256	.127
23	2.81	2.50	2.07	1.71	1.32	.858	.685	.532	.256	.127
24	2.80	2.49	2.06	1.71	1.32	.857	.685	.531	.256	.127
25	2.79	2.48	2.06	1.71	1.32	.856	.684	.531	.256	.127
26	2.78	2.48	2.06	1.71	1.32	.856	.684	.531	.256	.127
27	2.77	2.47	2.05	1.70	1.31	.855	.684	.531	.256	.127
28	2.76	2.47	2.05	1.70	1.31	.855	.683	.530	.256	.127
29	2.76	2.46	2.04	1.70	1.31	.854	.683	.530	.256	.127
30	2.75	2.46	2.04	1.70	1.31	.854	.683	.530	.256	.127
40	2.70	2.42	2.02	1.68	1.30	.851	.681	.529	.255	.126
60	2.66	2.39	2.00	1.67	1.30	.848	.679	.527	.254	.126
120	2.62	2.36	1.98	1.66	1.29	.845	.677	.526	.254	.126
∞	2.58	2.33	1.96	1.645	1.28	.842	.674	.524	.253	.126

Fuente: R. A. Fisher y F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research* (Tablas de estadísticas para la investigación biológica, agrícola y médica) (5a. edición), Tabla III, Oliver and Boyd Ltd., Edinburgh, con autorización de los autores y editores.

Apéndice IV

Valores percentiles (χ^2_p) correspondientes a la distribución ji cuadrada con ν grados de libertad (área sombreada = p)

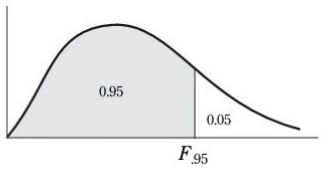


ν	$\chi^2_{.995}$	$\chi^2_{.99}$	$\chi^2_{.975}$	$\chi^2_{.95}$	$\chi^2_{.90}$	$\chi^2_{.75}$	$\chi^2_{.50}$	$\chi^2_{.25}$	$\chi^2_{.10}$	$\chi^2_{.05}$	$\chi^2_{.025}$	$\chi^2_{.01}$	$\chi^2_{.005}$
1	7.88	6.63	5.02	3.84	2.71	1.32	.455	.102	.0158	.0039	.0010	.0002	.0000
2	10.6	9.21	7.38	5.99	4.61	2.77	1.39	.575	.211	.103	.0506	.0201	.0100
3	12.8	11.3	9.35	7.81	6.25	4.11	2.37	1.21	.584	.352	.216	.115	.072
4	14.9	13.3	11.1	9.49	7.78	5.39	3.36	1.92	1.06	.711	.484	.297	.207
5	16.7	15.1	12.8	11.1	9.24	6.63	4.35	2.67	1.61	1.15	.831	.554	.412
6	18.5	16.8	14.4	12.6	10.6	7.84	5.35	3.45	2.20	1.64	1.24	.872	.676
7	20.3	18.5	16.0	14.1	12.0	9.04	6.35	4.25	2.83	2.17	1.69	1.24	.989
8	22.0	20.1	17.5	15.5	13.4	10.2	7.34	5.07	3.49	2.73	2.18	1.65	1.34
9	23.6	21.7	19.0	16.9	14.7	11.4	8.34	5.90	4.17	3.33	2.70	2.09	1.73
10	25.2	23.2	20.5	18.3	16.0	12.5	9.34	6.74	4.87	3.94	3.25	2.56	2.16
11	26.8	24.7	21.9	19.7	17.3	13.7	10.3	7.58	5.58	4.57	3.82	3.05	2.60
12	28.3	26.2	23.3	21.0	18.5	14.8	11.3	8.44	6.30	5.23	4.40	3.57	3.07
13	29.8	27.7	24.7	22.4	19.8	16.0	12.3	9.30	7.04	5.89	5.01	4.11	3.57
14	31.3	29.1	26.1	23.7	21.1	17.1	13.3	10.2	7.79	6.57	5.63	4.66	4.07
15	32.8	30.6	27.5	25.0	22.3	18.2	14.3	11.0	8.55	7.26	6.26	5.23	4.60
16	34.3	32.0	28.8	26.3	23.5	19.4	15.3	11.9	9.31	7.96	6.91	5.81	5.14
17	35.7	33.4	30.2	27.6	24.8	20.5	16.3	12.8	10.1	8.67	7.56	6.41	5.70
18	37.2	34.8	31.5	28.9	26.0	21.6	17.3	13.7	10.9	9.39	8.23	7.01	6.26
19	38.6	36.2	32.9	30.1	27.2	22.7	18.3	14.6	11.7	10.1	8.91	7.63	6.84
20	40.0	37.6	34.2	31.4	28.4	23.8	19.3	15.5	12.4	10.9	9.59	8.26	7.43
21	41.4	38.9	35.5	32.7	29.6	24.9	20.3	16.3	13.2	11.6	10.3	8.90	8.03
22	42.8	40.3	36.8	33.9	30.8	26.0	21.3	17.2	14.0	12.3	11.0	9.54	8.64
23	44.2	41.6	38.1	35.2	32.0	27.1	22.3	18.1	14.8	13.1	11.7	10.2	9.26
24	45.6	43.0	39.4	36.4	33.2	28.2	23.3	19.0	15.7	13.8	12.4	10.9	9.89
25	46.9	44.3	40.6	37.7	34.4	29.3	24.3	19.9	16.5	14.6	13.1	11.5	10.5
26	48.3	45.6	41.9	38.9	35.6	30.4	25.3	20.8	17.3	15.4	13.8	12.2	11.2
27	49.6	47.0	43.2	40.1	36.7	31.5	26.3	21.7	18.1	16.2	14.6	12.9	11.8
28	51.0	48.3	44.5	41.3	37.9	32.6	27.3	22.7	18.9	16.9	15.3	13.6	12.5
29	52.3	49.6	45.7	42.6	39.1	33.7	28.3	23.6	19.8	17.7	16.0	14.3	13.1
30	53.7	50.9	47.0	43.8	40.3	34.8	29.3	24.5	20.6	18.5	16.8	15.0	13.8
40	66.8	63.7	59.3	55.8	51.8	46.6	39.3	33.7	29.1	26.5	24.4	22.2	20.7
50	79.5	76.2	71.4	67.5	63.2	56.3	49.3	42.9	37.7	34.8	32.4	29.7	28.0
60	92.0	88.4	83.3	79.1	74.4	67.0	59.3	52.3	46.5	43.2	40.5	37.5	35.5
70	104.2	100.4	95.0	90.5	85.5	77.6	69.3	61.7	55.3	51.7	48.8	45.4	43.3
80	116.3	112.3	106.6	101.9	96.6	88.1	79.3	71.1	64.3	60.4	57.2	53.5	51.2
90	128.3	124.1	118.1	113.1	107.6	98.6	89.3	80.6	73.3	69.1	65.6	61.8	59.2
100	140.2	135.8	129.6	124.3	118.5	109.1	99.3	90.1	82.4	77.9	74.2	70.1	67.3

Fuente: Catherine M. Thompson, *Table of percentage points of the χ^2 distribution.* Biometrika, vol. 32 (1941) con autorización de autor y editor.

Apéndice V

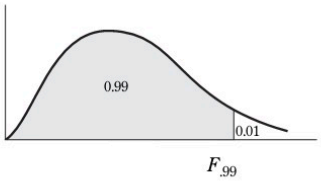
Valores del percentil 95 correspondientes a la distribución *F*
(ν_1 grados de libertad en el numerador)
(ν_2 grados de libertad en el denominador)



ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Fuente: E. S. Pearson y H. O. Hartley, *Biometrika Table for Statisticians*, vol. 2 (1972), tabla 5, página 178, con autorización.

Valores del percentil 99 correspondientes a la distribución *F*
(ν_1 grados de libertad en el numerador)
(ν_2 grados de libertad en el denominador)



ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	5000	5403	5625	5764	5859	5928	5981	6023	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.82	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Fuente: E. S. Pearson y H. O. Hartley, *Biometrika Table for Statisticians*, vol. 2 (1972), tabla 5, página 180, con autorización.

Logaritmos comunes con cuatro cifras decimales

N	0	1	2	3	4	5	6	7	8	9	Partes proporcionales 1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	16	19	21
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	3	4	5	6	7	8
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1	2	2	3	4	5	6	7	7
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7
N	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

Logaritmos comunes con cuatro cifras decimales (*continuación*)

N	0	1	2	3	4	5	6	7	8	9	Partes proporcionales 1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	1	2	3	4	4	5	6	7
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	1	2	3	4	4	5	6	7
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	5	5	6
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	3	3	4	4	5	6
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1	1	2	2	3	4	4	5	6
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1	1	2	2	3	4	4	5	6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	5
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	5
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1	2	2	3	4	4	5	5
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	5
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	5
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	0	1	1	2	2	3	3	4	4
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0	1	1	2	2	3	3	4	4
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0	1	1	2	2	3	3	4	4
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0	1	1	2	2	3	3	3	4
N	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9

Apéndice VIII

Valores de $e^{-\lambda}$

$(0 < \lambda < 1)$

λ	0	1	2	3	4	5	6	7	8	9
0.0	1.0000	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9231	.9139
0.1	.9048	.8958	.8869	.8781	.8694	.8607	.8521	.8437	.8353	.8270
0.2	.8187	.8106	.8025	.7945	.7866	.7788	.7711	.7634	.7558	.7483
0.3	.7408	.7334	.7261	.7189	.7118	.7047	.6977	.6907	.6839	.6771
0.4	.6703	.6636	.6570	.6505	.6440	.6376	.6313	.6250	.6188	.6126
0.5	.6065	.6005	.5945	.5886	.5827	.5770	.5712	.5655	.5599	.5543
0.6	.5488	.5434	.5379	.5326	.5273	.5220	.5169	.5117	.5066	.5016
0.7	.4966	.4916	.4868	.4819	.4771	.4724	.4677	.4630	.4584	.4538
0.8	.4493	.4449	.4404	.4360	.4317	.4274	.4232	.4190	.4148	.4107
0.9	.4066	.4025	.3985	.3946	.3906	.3867	.3829	.3791	.3753	.3716

$(\lambda = 1, 2, 3, \ldots, 10)$

λ	1	2	3	4	5	6	7	8	9	10
$e^{-\lambda}$	.36788	.13534	.04979	.01832	.006738	.002479	.000912	.000335	.000123	.000045

Nota: Para obtener valores de $e^{-\lambda}$ correspondientes a otros valores de λ, úsense las leyes de los exponentes.
Ejemplo: $e^{-3.48} = (e^{-3.00})(e^{-.48}) = (0.04979)(0.6188) = 0.03081$.

Apéndice IX

Números aleatorios

51772	74640	42331	29044	46621	62898	93582	04186	19640	87056
24033	23491	83587	06568	21960	21387	76105	10863	97453	90581
45939	60173	52078	25424	11645	55870	56974	37428	93507	94271
30586	02133	75797	45406	31041	86707	12973	17169	88116	42187
03585	79353	81938	82322	96799	85659	36081	50884	14070	74950
64937	03355	95863	20790	65304	55189	00745	65253	11822	15804
15630	64759	51135	98527	62586	41889	25439	88036	24034	67283
09448	56301	57683	30277	94623	85418	68829	06652	41982	49159
21631	91157	77331	60710	52290	16835	48653	71590	16159	14676
91097	17480	29414	06829	87843	28195	27279	47152	35683	47280
50532	25496	95652	42457	73547	76552	50020	24819	52984	76168
07136	40876	79971	54195	25708	51817	36732	72484	94923	75936
27989	64728	10744	08396	56242	90985	28868	99431	50995	20507
85184	73949	36601	46253	00477	25234	09908	36574	72139	70185
54398	21154	97810	36764	32869	11785	55261	59009	38714	38723
65544	34371	09591	07839	58892	92843	72828	91341	84821	63886
08263	65952	85762	64236	39238	18776	84303	99247	46149	03229
39817	67906	48236	16057	81812	15815	63700	85915	19219	45943
62257	04077	79443	95203	02479	30763	92486	54083	23631	05825
53298	90276	62545	21944	16530	03878	07516	95715	02526	33537

ÍNDICE

A
Abscisa, 4
Ajuste de curva, 316
  método a mano de, 318
  método de mínimos cuadrados de, 319
Ajuste de datos, 19 (*ver también* Ajuste de curva)
  empleando papel para probabilidad, 177
  mediante una distribución binomial, 195
  mediante una distribución de Poisson, 198
  mediante una distribución normal, 196
Aleatorio/a(s):
  errores, 403
  muestra, 203
  números, 203
  variable, 142
Aleatorización completa, 413
Análisis combinatorio, 146
Análisis de varianza, 362-401
  experimentos de dos factores usando, 407
  experimentos de un factor usando, 403
  modelo matemático para, 403
  objetivo de, 403
  tablas, 406
  usando cuadrados grecolatinos, 413
  usando cuadrados latinos, 413
Aproximación de curvas, ecuaciones de, 317
Aproximación de Stirling a $n!$, 146
Aproximación normal a la distribución binomial, 174
Áreas:
  en la distribución ji-cuadrada, 277
  en la distribución F, 279
  en la distribución normal, 173
  en la distribución t, 275
Artículo defectuoso, 487
Artículo no conforme, 486
Asintóticamente normal, 204
Atributos, correlación de, 298
Autocorrelación, 351

B
Base, 2
  de logaritmos comunes, 6
  de logaritmos naturales, 6
Bivariada:
  distribución normal, 351
  población, 351
  tabla o distribución de frecuencia, 366
Bloques aleatorizados, 413

C
Cálculos, 3
  reglas para, 3
  reglas para, empleando logaritmos, 7
Carta-C, 490
Carta de control de atributos, 481
Categorías, 37
Causas asignables, 480
Causas comunes, 480
Causas especiales, 480
Centro de gravedad, 320
Centroide, 320
Clase, 37 (*ver también* Intervalos de clase)
Clase modal, 43
Clasificación en un sentido, 403
Clasificaciones de dos sentidos, 407
Coeficiente cuartílico de asimetría, 125
Coeficiente cuartílico de dispersión relativa, 116
Coeficiente de asimetría de Pearson, 125
Coeficiente de correlación, 348
  de tablas de contingencia, 298
  fórmula producto momento para, 350
  para datos agrupados, 350
  rectas de regresión y, 351
  teoría del muestreo y, 351
Coeficiente de correlación de orden cero, 383
Coeficiente momento de curtosis, 125
Coeficiente momento de sesgo, 125
Coeficiente percentil de curtosis, 125
Coeficientes binomiales, 173
  triángulo de Pascal para, 180
Coeficientes de regresión parcial, 382
Combinaciones, 146
Comprobación de Charlier, 99
  para la media y la varianza, 99
  para momentos, 124

Conjunto nulo, 146
Constante, 1
Conteos o enumeraciones, 2
Contingencia, coeficiente de, 310
Coordenadas, rectangulares, 4
Corrección de Sheppard para momentos, 124
Corrección de Sheppard para varianza, 100
Corrección de Yates para continuidad, 297
Correlación, 345
  auto-, 351
  coeficiente de (*ver* Coeficiente de correlación)
  de atributos, 311
  espuria, 549
  lineal, 345
  medidas de, 346
  parcial, 385
  positiva y negativa, 345
  rango de, 450
  simple, 345
Correlación espuria, 349
Correlación múltiple, 382
Correlación parcial, 382
Correlación positiva, 346
Correlación simple, 382
Correlación sin sentido, 350
Correlación tetracórica, 299
Corridas, 449
Cuadrado latino ortogonal, 413
Cuadrados grecolatinos, 413
Cuadrados latinos, 413
  ortogonales, 413
Cuadrantes, 4
Cuadrática:
  curva, 317
  ecuación, 35
  función, 17
Cuantiles, 66
Cuartiles, 66
Curtosis, 125
  coeficiente momento de, 125
  coeficiente percentílico de, 125
  de una distribución binomial, 173
  de una distribución de Poisson, 176
  de una distribución normal, 125
Curva cuártica, 317
Curva cúbica, 317
Curva de frecuencia bimodal, 41
Curva de frecuencia multimodal, 41
Curva de Gompertz, 318
Curva de grado *n*, 317
Curva exponencial, 318
Curva geométrica, 318
Curva logística, 318
Curva normal, 173
  áreas bajo la, 173
  forma estándar de, 173
  ordenadas de, 187
Curvas de frecuencia, 41
  relativa, 41
  tipos de, 41
Curvas de frecuencia en forma de U, 41
Curvas de frecuencia sesgadas, 41
Curva simétrica o en forma de campana, 41

D

Datos:
  agrupados, 38
  dispersión o variación de, 89
  en bruto, 37
Datos agrupados, 38
Datos continuos, 1
  representación gráfica de, 57
Datos de atributos, 481
Datos discretos, 2
  representación gráfica de, 54
Datos en bruto, 76
Deciles, 66
  de datos agrupados, 87
  errores estándar para, 206
Decisiones estadísticas, 245
Defectos, 487
Desigualdades, 5
Desviación de la media aritmética, 63
Desviación estándar, 96
  de datos agrupados, 98
  de distribuciones muestrales, 204
  de una distribución de probabilidad, 155
  intervalo de confianza para, 230
  método abreviado de cálculo, 98
  método de codificación para, 96
  propiedades de, 98
  propiedad minimal de, 98
  relación con la desviación media, 100
  relación entre poblacional y muestral, 97
Desviación media, 95
  de la distribución normal, 174
  para datos agrupados, 96
Determinación:
  coeficiente de, 348
  múltiple, coeficiente de, 384
Determinación múltiple, coeficiente de, 384
Diagrama de Euler, 146
Diagrama de Venn (*ver* Diagrama de Euler)
Diagramas (*ver* Gráficas)
Dígitos o cifras significativas, 3
Diseño de experimentos, 412
Diseño experimental, 412
Dispersión, 87 (*ver también* Variación)
  absoluta, 95
  coeficiente de, 100
  medidas de, 95
  relativa, 100
Dispersión absoluta, 100
Dispersión del proceso, 282

Dispersión relativa o varianza, 95
Distribución acumulada "o más", 40
Distribución binomial, 172
  ajuste de datos, 195
  propiedades de, 172
  prueba de hipótesis usando, 250
  relación con la distribución de Poisson, 175
  relación con la distribución normal, 174
Distribución ji-cuadrada, 297 (*ver también* Ji-cuadrada)
  intervalos de confianza usando, 278
  pruebas de hipótesis y significancia, 294
Distribución de Bernoulli (*ver* Distribución binomial)
Distribución de Poisson, 175
  ajuste de datos mediante, 198
  propiedades de, 175
  relación con las distribuciones binomial y normal, 176
Distribución en forma de *J* inversa, 41
Distribuciones de frecuencia, 37
  acumulada, 40
  porcentual o relativa, 39
  reglas para la elaboración de, 38
Distribuciones de probabilidad acumulada, 143
Distribuciones de probabilidad continua, 143
Distribuciones de probabilidad discreta, 142
Distribuciones muestrales, 204
  de diferencias y sumas, 205
  de diversos estadísticos, 204
  de medias, 204
  de proporciones, 205
  experimentales, 208
Distribuciones unilaterales y bilaterales, 247
Distribución *F*, 279 (*ver también* Análisis de varianza)
Distribución multinomial, 177
Distribución normal, 173
  forma estándar de, 174
  proporciones de, 174
  relación con la distribución binomial, 174
  relación con la distribución de Poisson, 176
Distribución *t*, 275
Distribución unimodal, 64
Dominio de una variable, 1

E

Ecuaciones, 5
  cuadráticas, 35
  de curvas de aproximación, 317
  de regresión, 382
  equivalentes, 5
  miembros izquierdo y derecho de, 5
  simultáneas, 25, 34
  solución de, 5
  trasposición en, 26
Ecuaciones normales:
  para el plano de mínimos cuadrados, 321
  para la parábola de mínimos cuadrados, 320
  para la recta de mínimos cuadrados, 320
Ecuaciones simultáneas, 5
Ejes *X* y *Y* de un sistema de coordenadas rectangulares, 4
Eje *Y*, 4
Empates en la prueba *H* de Kruskal-Wallis, 448
  en la prueba *U* de Mann-Whitney, 447
Entrada de una tabla, 42
Enumeraciones, 2
Error de agrupación, 39
Error estándar de estimación, 348
  modificado, 348
Error probable, 230
Errores:
  de agrupación, 39
  de redondeo, 2
  probables, 230
Errores de redondeo, 2
Errores de redondeo acumulados, 2
Errores estándar de distribuciones muestrales, 206
Errores tipo I y tipo II, 246
Espacio de cuatro dimensiones, 385
Espacio muestral, 146
Esperanza matemática, 144
Estadística, 1
  deductiva o descriptiva, 1
  definición de, 1
  inductiva o inferencial, 1
Estadística deductiva, 1
Estadística descriptiva, 1
Estadística inductiva o inferencial, 1
Estadístico de prueba, 247
Estadístico *H*, 448
Estadístico muestral, 1
Estadístico *U*, 447
Estadísticos muestrales, 203
Estimación, 227
Estimaciones (*ver también* Estimación)
  eficientes e ineficientes, 228
  puntual y de intervalo, 228
  sesgadas y no sesgadas, 227
Estimaciones insesgadas, 227
Estimaciones por intervalo, 228
Estimaciones y estimadores eficientes, 228
Estimaciones y estimadores ineficientes, 228
Estimación puntual, 228
Estimación sesgada, 227
Evento compuesto, 143
Eventos, 140
  compuestos, 140
  dependientes, 140
  independientes, 140
  mutuamente excluyentes, 141
Eventos dependientes, 140
Eventos independientes, 140
Eventos mutuamente excluyentes, 141
Éxito, 139
Expansión multinomial, 177
Experimento de dos factores, 407
Experimento de un factor, 403
Exponente, 2

F
Factores de ponderación, 62
Factorial, 143
Fórmula de Spearman para la correlación por rangos, 450
Fórmula de Stirling para, 146
Fórmula o expansión binomial, 173
Fórmula para el interés compuesto, 85
Fórmula producto momento para el coeficiente de correlación, 350
Fracaso, 139
Frecuencia (*ver también* Frecuencia de clase)
  acumulada, 40
  modal, 41
  relativa, 40
Frecuencia acumulada, 40
Frecuencia de clase, 37
Frecuencia relativa, 39
  curvas, 39
  definición de probabilidad, 139
  distribución, 39
  tabla, 39
Frecuencias de celda, 496
Frecuencias esperadas o teóricas, 181
Frecuencias marginales, 294
Frecuencias observadas, 294
Frecuencias teóricas, 295
Fronteras de clase, inferior y superior, 38
Función, 4
  de distribución, 142
  de frecuencia, 142
  de probabilidad, 139
  lineal, 318
  multivaluada, 4
  univaluada, 4
Función de densidad, 144
Función de distribución, 142
Función de frecuencia, 142
Función de probabilidad, 142
Funciones univaluadas, 4

G
Gosset, 276
Grados de libertad, 276
Gráfica, 5
  de barras, 18
  de pastel, 5
Gráfica cero, 489
Gráfica circular (*ver* Gráfica o carta de pastel)
Gráfica Cusum, 490
Gráfica de efectos principales, 429
Gráfica de interacción, 429
Gráfica de medianas, 489
Gráfica de Pareto, 504
Gráfica EWMA, 490
Gráfica *NP*, 487
Gráfica o carta de pastel, 5
Gráfica-*P*, 487
Gráfica *U*, 490
Gráficas de control, 480
Gráficas individuales, 489
Gráficas o cartas de barras, parte componente, 4
Gráficas *X* barra y *R*, 481
Gran media, 404

H
Hipérbola, 318
Hiperplano, 385
Hipótesis, 245
  alternativa, 245
  nula, 245
Hipótesis alternativa, 245
Hipótesis nula, 245
Histogramas, 39
  cálculo de medianas para, 64
  de frecuencia porcentual o relativa, 39
  de probabilidad, 153
Hoja de conteo, 39

I
Identidad, 5
Índice CP, 485
Índice CPK, 485
Índice de capacidad del proceso, 485
Índice de capacidad inferior, 485
Índice de capacidad superior, 485
Interacción, 411
Interés compuesto, 85
Interpolación, 7
Intersección, 319
Intersección con el eje *X*, 283, 289-291
Intersección de conjuntos, 146
Intervalo de confianza:
  en correlación y regresión, 374
  para desviaciones estándar, 230
  para diferencias y sumas, 230
  para medias, 229
  para proporciones, 229
  usando la distribución ji-cuadrada, 278
  usando la distribución normal, 229-230
  usando la distribución *t*, 276
Intervalos de clase, 38
  abierto, 38
  amplitud o tamaño de, 38
  desiguales, 51
  mediano, 64
  modal, 43

J
Ji-cuadrada, 294
  correlación de Yates para, 297
  definición de, 294
  fórmulas para, en tablas de contingencia, 297
  para bondad de ajuste, 295
  propiedad aditiva de, 299
  prueba, 295

L
Leptocúrtica, 125
Límite inferior de control, 480
Límite inferior de especificación, 484
Límite superior de control, 480
Límite superior de especificación, 484
Límites de clase, 38
  inferior y superior, 38
  verdaderos, 38
Límites de confianza, 229
Límites de control, 480
Límites de especificación, 484
Línea central, 480
Línea recta, 317
  ecuación de, 317
  mínimos cuadrados, 318
  pendiente de, 318
  regresión, 321
Logaritmos, 6
  base de, 6
  cálculos usando, 7
  comunes, 6
  naturales, 6
Logaritmos comunes, 6
Logaritmos naturales, base de, 6
Longitud, tamaño o amplitud de clase, 38

M
Marca de clase, 38
Media aritmética, 61
  calculada a partir de datos agrupados, 63
  comprobación de Charlier para, 99
  de medias aritméticas, 63
  de una población y de una muestra, 144
  distribuciones de probabilidad, 142
  efecto de valores extremos (o atípicos) en, 71
  método de codificación para calcularla, 63
  método largo y método abreviado para calcularla, 63
  ponderada, 62
  propiedades de la, 63
  relación con la mediana y la moda, 64
  relación con las medias geométrica y armónica, 66
  supuesta o adivinada, 63
Media armónica, 65
  ponderada, 86
  relación con las medias aritmética y geométrica, 66
Media del grupo, 404
Media geométrica, 65
  de datos agrupados, 83
  idoneidad para promedios de cocientes, 84
  ponderada, 84
  relación con las medias aritmética y armónica, 66
Mediana, 64
  calculada a partir de un histograma o de una ojiva porcentual, 64
  de datos agrupados, 64
  efecto de valores atípicos sobre la, 78
  relación con la media aritmética y con la moda, 64
Medias de renglón, 403
Medias de tratamiento, 345
Mediciones, 2
Medidas de tendencia central, 61
Mejor estimación, 228
Mesocúrtica, 125
Método a mano para el ajuste de curvas, 318
Método de conteo en la prueba *U* de Mann-Whitney, 447
Métodos de codificación, 63
  para el coeficiente de correlación, 350
  para el momento, 124
  para la desviación estándar, 98
  para la media, 63
Mínimos cuadrados:
  curva, 319
  parábola, 320
  plano, 321
  recta, 319
Moda, 64
  de datos agrupados, 64
  fórmulas para, 64
  relación con la media y la mediana, 65
Modelo o distribución teórica, 177
Momentos, 123
  adimensionales, 124
  correcciones de Sheppard para, 124
  definición de, 123
  método de codificación para el cálculo de, 124
  para datos agrupados, 123
  relaciones entre, 124
  verificación de Charlier para el cálculo de, 124
Momentos adimensionales, 124
Muestra, 1
Muestreo:
  con reemplazo, 204
  sin reemplazo, 204

N
Nivel de significancia, 246
Niveles de confianza, tablas de, 229
No aleatoria, 449
No lineal:
  correlación y regresión, 346
  ecuaciones, reducibles a forma lineal, 320
  regresión múltiple, 382
  relación entre variables, 316
Notación científica, 2
Número muestral, 213

O
Ojivas, 40
  deciles, percentiles y cuartiles obtenidos de, 87
  mediana obtenida de, 78
  "menos de", 52
  "o más", 52
  porcentual, 53
  suavizada, 53

Ordenaciones, 37
Ordenadas, 4
Origen, 5

P
Papel para gráficas
de probabilístico, 177
log-log, 339
semilogarítmico, 318
Papel semilog, 318
Parábola, 320
Parámetros, estimación de, 227
Parámetros poblacionales, 228
Partes por millón (ppm), 484
Patrón cíclico en pruebas de corridas, 449
Pendiente de una recta, 318
Percentiles, 66
Permutaciones, 145
Plano, 4
Plano *XY*, 4
Platicúrtica, 125
Población, 1
Polígonos de frecuencia, 39
porcentuales o relativos, 39
suavizados, 41
Polinomios, 318
Ponderada:
media aritmética, 62
media armónica, 85
media geométrica, 83
Porcentual:
distribución, 39
distribuciones acumuladas, 39
frecuencia acumulada, 39
histograma, 38
ojivas, 39
Probabilidad, 139
análisis combinatorio y, 143
axiomática, 140
condicional, 140
definición clásica de, 139
definición de, mediante frecuencias relativas, 140
distribuciones de, 142
empírica, 140
reglas fundamentales de, 146
relación con la teoría de conjuntos, 146
Probabilidad condicional, 140
Probabilidad empírica, 139
Progresión aritmética:
momentos de, 136
varianza de, 120
Promedio, 62
Proporción de no conformes, 484
Proporciones, 205
distribución muestral de, 205
intervalo de confianza para, 229
pruebas de hipótesis para, 245
Prueba de bondad de ajuste, 177 (*véase también* Ajuste de datos)
Prueba de dos lados o de dos colas, 247
Prueba de homogeneidad de la varianza, 285
Prueba de los signos, 446
Prueba de normalidad, 196
Prueba *H* de Kruskal-Wallis, 448
Pruebas:
de hipótesis y de significancia, 245
en las que interviene la distribución binomial, 250
en las que interviene la distribución normal, 246
para causas especiales, 484
para diferencias de medias y proporciones, 249
para medias y proporciones, 249
relacionadas con correlación y regresión, 352
Pruebas no paramétricas, 446
para correlación, 450
prueba de corridas, 449
prueba de los signos, 449
prueba *H* de Kruskal-Wallis, 448
prueba *U* de Mann-Whitney, 447
Prueba *U* de Mann Whitney, 454
Punto cero, 4
Puntuaciones estándar, 101
Puntuación o estadístico *t*, 275

Q
Quintiles, 87

R
Raíz cuadrada media, 66
Rango, 95
intercuartílico, 96
percentil 10-90, 96
semiintercuartílico, 96
Rango, coeficiente de correlación por, 450
Rango intercuartílico, 96
semi-, 96
Rango percentil, 96
Rango percentílico 10-90, 96
Rango semiintercuartílico, 96
Redondeo de datos, 2
Región crítica, 247
Región de aceptación, 247 (*ver también* Hipótesis)
Reglas de decisión, 246 (*ver también* Decisiones estadísticas)
Regresión, 321
curva de, 321
múltiple, 345
plano, 321
recta, 321
simple, 343
superficie, 322
teoría muestral de, 352
Relación empírica entre media, mediana y moda, 64
Relación empírica entre medidas de dispersión, 100
Residual, 319

S
Sesgo, 41
  coeficiente cuartílico de, 125
  coeficiente de sesgo percentílico 10-90, 125
  coeficientes de Pearson de, 125
  distribución binomial, 172
  distribución de Poisson, 175
  distribución normal, 173
  negativo, 41
Signos de desigualdad, 5
Solución de ecuaciones, 5
Subgrupos, 418
Subíndices, 61
Sumatoria, 61

T
Tabla, 4
Tabla de correlación, 350
Tablas de contingencia, 296
  coeficiente de correlación de, 298
  fórmulas para ji-cuadrada en, 298
Tablas de frecuencia (*ver también* Distribuciones de frecuencia)
  acumulada, 40
  relativa, 40
Teorema del límite central, 204
Teorema o regla de Bayes, 170
Teoría de las muestras pequeñas, 275
Teoría del muestreo, 203
  de correlación, 351
  de regresión, 352
  muestras grandes, 207
  uso de, en estimación, 227
  uso de, en pruebas de hipótesis y de significancia, 245
Teoría exacta del muestreo o teoría de las muestras pequeñas, 181
Tratamiento, 345
Triángulo de Pascal, 180

U
Unidad de inspección, 489
Unión de conjuntos, 146

V
Valor absoluto, 95
Valores críticos, 248
Valores críticos (o coeficientes de confianza), 228
valor-*p*, 245
Variable, 1
  continua, 1
  datos, 480
  dependiente, 4
  discreta, 1
  distribuida normalmente, 173
  dominio de, 1
  estandarizada, 101
  gráfica de control, 480
  independiente, 4
  relación entre, 316 (*ver también* Ajuste de curva; Correlación; Regresión)
Variable continua, 1
Variable dependiente, 4
  cambio de, en ecuaciones de regresión, 284
Variable discreta, 1
Variable estandarizada, 101
Variable independiente, 4
Variación, 95 (*ver también* Dispersión)
  coeficiente cuartílico de, 116
  coeficiente de, 100
  explicada y no explicada, 348
  residual, 408
  total, 348
Variación explicada, 348
Variación no explicada, 348
Variación residual, 409
Variación total, 348
Varianza, 97 (*ver también* Desviación estándar)
  combinada o conjunta, 98
  comprobación de Charlier para, 99
  corrección de Sheppard para, 100
  de una distribución de probabilidad, 204
  muestral modificada, 227
  relación entre poblacional y muestral, 14